The genetic variation of cultivated flax (*Linum usitatissimum* L.) and the role of its wild ancestor (*Linum bienne* Mill.) in its evolution.

by

Rafał Marek Gutaker

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Warwick

School of Life Sciences

October 2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to dedicate this thesis to my closest family and friends; who have been always helpful and supportive. They created the environment, in which I could pursue my dreams of completing my PhD degree, were the source of motivation, and supported me in difficult times. I would like to thank my supervisor Robin Allaby who has been an inspiration into what it means to be creative and never hesitate to experiment. I am also indebted to my co-supervisor Dr. Sabine Karg.

I would also like to acknowledge all the people who helped me during the material collection. I am indebted to researchers who helped me in organizing the collection expedition (Prof. Toni Nikolic, Prof. Arne Strid, Dr. Siyka Angelova and Dr. Zlatko Satovic). I would like to say thank you to people who participated in my expedition (Ewa Samorzewska, Agnieszka Czarnocka-Cieciura and Mariusz Czarnocka-Cieciura). I would also like to acknowledge researchers who invited me to sample historic samples (Dr. Stephen Harris, Dr. Mark Nesbitt, Dr. Maja Graniszewska and Dr. Mark Spencer). I acknowledge Dr. Nina Brutch for contributing historic landraces. Finally, I would like to thank Dr. Axel Diderichsen and Dr. Yong-Bi Fu for the biggest contribution to the plant material used in this work and their continuous support and interest in my studies.

A Big thank you to the people who helped me with my experiments and analyses. I would like to especially thank Dr. Peter Beerli, Dr. John Davey and Dr. Julian Catchen for their support with developing and utilizing genomic data and Dr. Andrea Massiah and Dr. Tiziana Sgamma for their help in the growing experiment. I feel I should thank the Wikipedia, which helped me recalling long-forgotten facts.

Finally, I would like to thank my group mates who highlighted every day at work with their positive attitude. I know I can always count on them in my personal and academic life. Andrew Clarke, James Kitchen, Roselyn Ware, Oliver Smith and Sarah Palmer – thank you!

DECLARATIONS

This thesis is submitted to the University of Warwick in partial fulfilment of the requirements for admission to the degree of Doctor of Philosophy. The work presented here is my own, except where specifically stated otherwise and was performed in the School of Life Sciences at the University of Warwick under supervision of Doctor Robin G. Allaby during the period October 2010 to July 2014. The sampling of historic plant material was carried out at the University of Oxford Herbarium, Natural History Museum London British collection, Natural History Museum London International collection, Kew Gardens Economic Botany collection and the University of Warsaw Herbarium. Sampling of the wild populations of *Linum bienne* took place during the field expedition to the Balkan countries in July 2011. The modern cultivars of flax together with the Turkish accessions of pale flax were contributed by Plant Genetic Resources Canada. The landraces of cultivated flax were contributed by Vavilov Institute in Saint Petersburg, Russia. All the laboratory experiments and bioinformatic analyses were conducted in the School of Life Sciences at the University of Warwick. This work was presented at the Molecular and Genome Evolution Symposium in Manchester in May 2013 and at the Society for Molecular Biology and Evolution Meeting in Chicago in July 2013. This work has not yet been published in the form of scientific article.

SUMMARY

The aim of this thesis is to investigate the evolution of cultivated flax (*Linum usitatissimum* L.). Specifically, it focuses on the adaptation of flax to European latitudes during the spread of agriculture. The two main objectives of this study are to determine 1) the role of hybridization with the flax wild relative (pale flax; *Linum bienne* Mill.) in the adaptation to northerly latitudes and 2) the link between the adaptation and the emergence of fibre flax varieties.

Flowering time genes play an important role in adaptation to northerly latitudes. Homologs of three flowering genes from *Arabidopsis thaliana* (*LFY, TFL1* and *TOC1*) were identified and sequenced in flax. This included seven flax paralogs of *TFL1*, which shared gene structure, exonic sequences, conservation of amino acids in the active site, and expression patterns with the phosphatidylethanolamine binding protein family that includes TFL1. Sixteen different haplotypes of *LuTFL1* were discovered amongst sequence data derived from a diversity collection of flax accessions. Haplotypes in cluster I are associated with samples from Southern Europe and the Near East, while those in cluster III are associated with Central and Northern Europe. The *LuTFL1* gene was under selection or selective sweep that occurred in cultivated flax. Haplotype I was inherited from pale flax during domestication, while haplotype III was inherited post-domestication following hybridization with pale flax from around the Bosphorus. Within sequence data of *LuTFL2*, ten haplotypes were discovered and these too were correlated with latitude. Contrary to the previous case however, there is no evidence of selection in *LuTFL2*. *LuTFL1* and *LuTFL2* might be in linkage disequilibrium.

A collection of RADseq markers was developed to investigate the genetic structure of flax and measure the genetic migration between wild and cultivated species. Of all the RADseq markers, 219 were characterized with single SNP and used for further analyses. The genetic diversity of cultivated flax was found to be similar to that of pale flax. The population structure inferred from SNP data showed genetic separation of dehiscent and non-dehiscent varieties, but no distinction between intermediate, fibre and oil varieties of cultivated flax. The population is structured from south to north within landraces, intermediate and fibre varieties. There is strong evidence for admixture from pale flax to cultivated flax, especially from the populations in the area of domestication. High levels of genetic migration are shown from all populations of pale flax towards both the southern and northern cultivated flax. RADseq SNP alleles that are associated with northerly latitudes in cultivated flax were introduced through post-domestication hybridization with pale flax of various geographic origin.

These combined molecular results indicate that pale flax contributed to the adaptation of cultivated flax in the European climate through post-domestication gene flow. Considering archaeological and phenotypic data from the literature, the adaptation to northerly latitude might have led to the emergence of fibre varieties. This inference is supported by a custom-made programme called *PGROWTH*, which models the impact of *FT* and *TFL1* gene expression on flax architecture. These two genes play an important role in plant development and therefore might have an impact on flax architecture. This study suggests the need of carrying out transgenic, functional experiments with *LuTFL1* constructs in the future.

LIST OF COMMON ABBREVIATIONS

A	adenine
AFLP	amplified fragment length polymorphism
BCE	before common era
BLAST	Basic Local Alignment Search Tool
bp	base pairs
C	cytosine
CDs	coding sequence
CF	cultivated flax
- CF D	dehiscent variety of CF
- CF I	intermediate variety of CF
- CF L	landrace of CF
DDJB	DNA Data Bank of Japan
DEPC	diethylpyrocarbonate
DNA	deoxyribonucleic acid
- cDNA	complementary DNA
- cpDNA	chloroplast DNA
EST	expressed sequence tag
ExPASy	Expert Protein Analysis System
F_{ST}	fixation index
G	guanine
GBS	genotyping by sequencing
GNS	Geographic Names Server
H***	historic accession
IRAP	inter-transposon amplified polymorphism
ISSR	inter-simple sequence repeats
JTT	Jones, Taylor and Thornton model
LBK	Linear Pottery culture
LD	long days
LDE	linkage disequilibrium
MADS	*MCM1, AGAMOUS, DEFICIENS, SRF*
MCMC	Markov chain Monte-Carlo
ML	maximum likelihood

M***	modern accession
NA	not applicable
NCBI	National Centre for Biotechnology Information
NEB	New England Biolab
N_{em}	effective number of immigrants per generation
NGS	next generation sequencing
ORF	open reading frame
PCA	principle component analysis
PCR	polymerase chain reaction
PEBP	phosphatidylethanolamine binding protein
PF	pale flax
PGRC	Plant Genetic Resources Canada
QTL	quantitative trait locus
RAD	restriction site associated DNA
- RADseq	RAD sequencing
RAPD	random amplified polymorphic DNA
RFLP	restriction length polymorphism
RNA	ribonucleic acid
- miRNA	micro RNA
SD	short days
SNP	single nucleotide polymorphism
sp.	species
ssp.	subspecies
SSR	short sequence repeat
Θ	population mutation rates
T	thymine
TAIR	The Arabidopsis Information Resources
UPGMA	unweighted pair group method with arithmetic mean
UV	ultraviolet
W***	wild accession

LIST OF GENE NAME ABBREVIATIONS

ADG1	*ADP GLUCOSE PYROPHOSPHORYLASE 1*
AP1	*APETALA1*
ATC	*Arabidopsis thaliana CENTRORADIALIS*
BFT	*BROTHER OF FT AND TFL1*
CCA1	*CIRCADIAN CLOCK ASSOCIATED 1*
CEN	*CENTRORADIALIS*
CO	*CONSTANS*
ELF3	*EARLY FLOWERING 3*
FD	*FLOWERING LOCUS D*
FLC	*FLOWERING LOCUS C*
FRI	*FRIGIDA*
FT	*FLOWERING LOCUS T*
GAI	*GIBBERILLIC ACID INSENSITIVE*
GAPDH	*GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE*
GI	*GIGANTEA*
HOS1	*HIGH EXPRESSION OF OSMOTICALLY RESPONSIVE GENES 1*
HR	*HIGH RESPONSE TO PHOTOPERIOD*
ITS	*INTERNAL TRANSCRIBED SPACER*
LFY	*LEAFY*
LHY	*LATE ELONGATED HYPOCOTYL*
MFT	*MOTHER OF FT AND TFL*
PHYA	*PHYTOCHROME A*
PPD1	*PHOTOPERIOD DEPENDENT 1*
SAD2	*STEAROYL-ACP DESATURASE 2*
SFT	*SINGLE FLOWER TRUSS*
SOC1	*SUPPRESOR OF OVEREXPRESSION OF CONSTANS 1*
SP	*SELF PRUNING*
SVP	*SHORT VEGETATIVE PHASE*
TFL1	*TERMINAL FLOWERING LOCUS 1*
TOC1	*TIMING OF CAB 1*
TSF	*TWIN SISTER OF FT*
VRN1	*VERNALIZATION 1*

VRN2	*VERNALIZATION 2*
VRN3	*VERNALIZATION 3*

LIST OF GENE PREFIX ABBREVIATIONS

Bv	*Beta vulgaris*
Gm	*Glycine max*
Ha	*Helianthus annuus*
Hv	*Hordeum vulgare*
Lu	*Linum usitatissimum*
Mt	*Medicago tranculata*
Pn	*Populus nigra*
Pt	*Populus trichocarpa*
R	*Rosa* sp.

CHAPTER 1: GENERAL INTRODUCTION

1.1 OVERVIEW

Cultivated flax is an annual or biennial herbaceous plant that belongs to family Linaceae. It was domesticated around 12,000 years ago in the Near East from pale flax. From there, cultivated flax was spread with early farmers to Europe and Asia. The Near Eastern flax varieties had to adapt to climate at the northerly latitudes in order to be cultivated successfully in Central and Northern Europe.

Cultivated flax varieties which are grown for seeds are morphologically distinct from varieties that are grown for fibre. The archaeological evidence suggests that adaptation of cultivated flax to northerly latitudes might have been linked to a change in morphology, which in turn potentially triggered the emergence of fibre varieties. Cultivated flax is unusual in the geographic distribution of its wild relative (pale flax), which stretches far north from the domestication area. Hence it is hypothesised that northern pale flax populations might have contributed to an adaptation of cultivated flax to the new climate during its early spread and therefore be one of the factors leading to the emergence of the specialized fibre varieties.

In crops, adaptation to new climate is associated with changes in flowering time. There is little known about molecular mechanism of flowering in flax, but we can assume that many key genes are conserved between flax, poplars and *A. thaliana*. Genotyping markers associated with floral genes are a good starting point for investigating the latitudinal adaptation of flax. Genotyping neutral markers on the other hand are essential to investigate genetic population structure and migrations in flax subpopulations.

1.2 BACKGROUND

1.2.1 Taxonomy and morphology of cultivated flax

Cultivated flax (*Linum usitatissimum* L.) is a species that belongs to the family Linaceae of the order Malphigiales, group Fabidae, clade Rosids (Bremer *et al.* 2009). The family Linaceae is characterised by: 1) leaves - always simple; with arrangement varying from alternate to opposite or whorled, 2) flowers which are hermaphroditic, actinomorphic, pentameric, or very rarely tetrameric, blue or yellow, rarely red, white, or pink, and some are heterostylous and 3) fruit type is capsule with an average of 6-10 seeds. Within Linaceae there are two distinguishable subfamilies: Linoideae and Hugonioideae, together comprised of the 13 genera. *Linum* is the largest genus; it comprises of 187 species (McDill & Simpson 2011). Members of the genus *Linum* include herbaceous annuals, biennials and perennials that are native to temperate and subtropical climates.

A basis for distinguishing the five main sections of *Linum* was given in 19th century (Planchon 1848), which was developed later in the 20th century (Winkler 1931). Sections *Linopsis*, *Syllinum*, *Cathartolinum*, *Dasylinum* and *Linum* were later framed into a phylogenetic tree based on the chloroplast markers and ITS sequences (McDill *et al.* 2009). The molecular phylogeny of the section *Linum* was congruent with earlier morphological classification, with only *L. stelleroides* being excluded from an otherwise well-supported, monophyletic clade. Section *Linum* groups Eurasian species, including cultivated flax *L. usitatissimum*, pale flax *L. bienne*, *L. decumbens*, *L. grandiflorum* and *L. narbonese* with North American *L. lewisii* and Australian *L. marginale*. These species share the characteristic traits of lanceolate, nondecurrent leaves without stipular glands and flowers with bright blue petals.

Cultivated flax is an annual or biennial erect plant, usually with a single stem. Its leaves are linear, 10-45 mm long and three-nerved. Its flowers are characterized by the broadly elliptic sepals and blue petals that are 10-15 mm long (Davis 1988). The inflorescence is usually identified as raceme or panicle with alternate pedicels. After maturation the 7-8 mm wide capsule is indehiscent or subdehiscent containing brown, yellow or olive coloured seeds. *L. bienne* is similar to cultivated flax in many respects, however, there are some differences. Among the most notable differences is the number of stems, which in pale flax reach tens where they are procumbent,

ascending or erect, while in cultivated flax it is usually single. Research on the morphological diversity within individuals of these two species revealed that they differ greatly in the stem height, size of capsules, length of petals and weight of seeds, and all these traits are significantly lower in *L. bienne* (Diederichsen & Hammer 1995). The same report showed huge differences in phenology of the both species; cultivated flax flowers much earlier and for shorter period of time than pale flax, this could be a result of the annual growth regime of summer varieties of cultivated flax. Furthermore, pale flax capsules open after maturation (dehiscence), whereas cultivated flax capsules remain closed (indehiscence).

1.2.2 Domestication of cultivated flax

Domestication is an evolutionary process in which domesticated organism adapts to human environment. It adapts through changes in morphology and phenology, which often lead to diversification and emergence of new species. The domestication of crop plants occurred during cultivation and selection of wild plants by early farmers. Traits that were selected for in wild plants under domestication are collectively called 'the domestication syndrome' (Hammer 1984). The domestication syndrome includes both the morphological and phenological changes. Examples of the former are the loss of a natural seed dispersal mechanism (non shattering, indehiscence) and increased seed size. The phenology changes include: a loss of dormancy, reduced vernalization requirements and shift in photoperiodicity that triggers flowering (Fuller & Allaby 2009). Interestingly, domestication syndrome traits often boost the process of species divergence. Changes in phenology have major impact on species divergence through temporal isolation of reproductive stage. Furthermore, change from dehiscence to indehiscence restricts propagation of hybrids between wild and cultivated plants. Indehiscent, wild plants will not shed their seeds at all while seeds from dehiscent, cultivated plants will not be harvested and sown in the next season. As a consequence, we often observe partial or full speciation of domesticates from their wild relatives. Although ancestry of many crop plants is not obvious, one can identify their closest wild relative by using morphological, cytological and molecular markers.

There is the strong evidence that cultivated flax was domesticated from pale flax. Although they differ greatly in seed size, branching pattern, dehiscence and

phenology (Diederichsen & Hammer 1995), the changes in these traits are usually recognized as a part of the domestication syndrome. A very close relationship between these two species is suggested by the same number of chromosomes (2n =30), both being homostylous and self-fertilizing (Zohary & Hopf 2000), hence in the past they were considered as the subspecies of the same species (Hammer 1986). Moreover, the cytological studies indicate that karyotype of the two species differs only by one translocation (Gill & Yermanos 1967; Yermanos & Gill 1969). There is also strong molecular evidence for a close relationship of *L. bienne* and *L. usitatissimum*, these two species are sister taxa in a phylogenetic tree based on both chloroplast and ITS markers (McDill *et al.* 2009). Additionally, pale flax is closely associated with cultivated flax in genetic networks of cultivated flax populations in studies based on the *SAD2* locus (Allaby *et al.* 2005), cpDNA markers (Fu & Allaby 2010), EST-derived SSR markers (Fu & Peterson 2010), ISSR (Uysal *et al.* 2010) and IRAP (Smykal *et al.* 2011). The second phylogenetically closest wild relative – *L. decumbens* (McDill *et al*. 2009) is 36 mutations away from pale/cultivated flax cluster in the distance network based on cpDNA data (Fu & Allaby 2010). Thus, all the evidence confirms that pale flax is the closest wild relative of cultivated flax.

Archaeological evidence suggest that *L. usitatissimum* was domesticated in the Near East around 12.000 years ago and reached Europe around 5.000 BCE. The earliest record of pale flax in the Near East are seeds dated to 9.000–8.000 years BCE, excavated in the settlement Tel Abu Hureyra in Syria (Hillman 1975). Cultivated flax, which can be recognized by a significantly increased size of seeds, is recorded from 7.000 BCE in Tell Ramad, Syria (Van Zeist & Bakker-Heeres 1975). It was present in the archaeological sites in the Southern Levant, Anatolia and Eastern Fertile Crescent throughout 10^{th} to 7^{th} millennium BCE (Fuller *et al.* 2012). According to Zohary and Hopf (2000) flax was the earliest oil and fibre crop, and it was domesticated in the Near East as part of the 'Neolithic package' (which also includes einkorn, emmer, barley, lentil, chickpea and pea). The most archaeological evidence consists of well-preserved seeds because fibrous stems are much more susceptible to degradation. The earliest remains of flax textiles in the area of domestication were dated to 7th millennium BCE (Zohary & Hopf 2000). Flax appears in the archaeological sites of the Near East the earliest and therefore it was

inferred that it was its original domestication area. From there, flax was probably introduced to other geographic regions of Asia and into Europe.

1.2.3 The spread of agriculture to Europe

The common understanding of the origins of agriculture is that it was first adopted as life style in the Near East and then spread as a novelty to Europe and Asia. The evidence for all of the seven founder crops that were domesticated in this region (flax, einkorn, emmer, barley, lentil, chickpea and pea) strongly supports this notion (Zohary & Hopf 2000). It is believed that early farmers slowly migrated or diffused to Europe bringing with them crop plants and related agrarian technology. There are multiple ways to attempt reconstruction of the early farmers movements.

The spread of the agriculture to Europe was assessed with the population genetics approach and use of haplotype data to infer about the human migrations. These analyses are based on the assumption that local European hunter-gatherers differentiated genetically from Near Eastern populations, which changed their lifestyle to settled farming and migrated into Europe. It has been shown that there is genetic discontinuity between the late hunter-gatherers and the early farmers in Central Europe (Bramanti *et al.* 2009) and there is the evidence that these farmers show a genetic affinity with present-day populations from the Near East and Anatolia (Haak *et al.* 2010). When genomic DNA of 5000 year old hunter-gatherers and farmers from Scandinavia was compared it turned out that the latter bears similarity to modern-day South-eastern Europeans rather than Scandinavians (Skoglund *et al.* 2012). This result suggests that agriculture came to Scandinavia from the Southern Europe. Similar analysis of the northern Spanish archaeological specimens revealed that the spread of farmers fits a random dispersion model and it is hypothesized to be a result of maritime colonization (Hervella *et al.* 2012). Altogether, these data present evidence for the spread of early farmers from the Near East through the Balkans both to Central/Northern Europe and along the coast of Mediterranean Sea.

The spread of agriculture was reconstructed from the results of phylogeographic analyses that included historic varieties of crop plants. The usefulness of modern cultivars for this purpose is debatable. The Green Revolution and globalisation resulted in unification and diffusion of geography-specific signature of genetic information within cultivars in the 20-21st century. This signature is thought to be

well preserved in extant landraces. These landraces however, are usually scarce and provide limited coverage of Europe and Asia. Only historic material, which is stored in herbaria or seed museums, contain geography-specific signal for better phylogeographic analysis and allow reconstructing more dated movements (Jones *et al.* 2008; Lister *et al.* 2010). For example the investigation of the historic European barley for its *PPD-H1* gene revealed the latitudinal cline in crop distribution (Lister *et al.* 2009). The data supported by the SSR marker analysis provided evidence suggesting that the adaptation to climate of some areas, such as Alpine region, Britain and Scandinavia had an impact on timing of the agricultural spread (Jones *et al.* 2012). Based on the cpDNA of a forage crop – *Lolium perenne* two routes of the agriculture spread were reconstructed: one around the Mediterranean Sea and the other leading to the Northern Europe (Balfourier *et al.* 2000). However this pattern was not clear for another forage crop – *Festuca pratensis* (Fjellheim et al. 2006). The studies described above show that phylogeography of most of crop plants could be used in tracing the spread of early agriculture in Europe and with use of historic samples could give more details about spatio-temporal character of this movement.

Another more direct way of tracing the spread of agriculture is by comparing the presence of crop plant remains or associated artefacts such as pottery in dated archaeological layers throughout Asia and Europe. Overall comparison of presence of various crop and weed species in different parts of Europe by means of phylogenetic analysis yielded controversial results (Coward *et al.* 2008). On the one hand it shows clear relationship between the ancestral plant economies from the Near East and more derived agricultural assemblages in Greece, Former Yugoslavia, Italy and Spain. On the other hand, it places Bulgaria in a distant branch from the Anatolia and relates it more to Central European plant economies (Coward *et al.* 2008). It has also been reported that the diversity of crop plants used in Neolithic Central Europe is significantly lower than in other agricultural regions possibly due to difficulties in adaptation to a northern climate (Concolly *et al.* 2008). Based on the distribution of cereal grains in the archaeological sites in Europe it was inferred that there were two major routes of early farmers spread: along Mediterranean Sea and along Danube river leading further in a north-easterly direction. On the Mediterranean track cereal cultivation spread to Italy and Spain about 5750 years BCE. On the alternative route it reached the Balkans around 5950 years BCE where it stopped for around 800 years

and then continued spreading rapidly to North-west Europe (Jones *et al.* 2007). It might be the case that the spread of agriculture along the Mediterranean Sea was undisturbed while the movement to Central Europe and towards north was slowed down by necessity of adapting to the new climate. Central Europe agriculture is believed to have spread together with Linear Pottery Ceramic Culture (LBK). The earliest LBK artefacts were found in the Hungarian Plain around 5700 BCE; from there the influence of this culture is believed to spread East, North and West. Both archaeological and genetic evidence robustly support existence of the two main routes of spread.

There is limited data available for tracing the spread of cultivated flax. Only archaeobotanical and archaeological data allow investigating flax early movements. To the best of the authors knowledge there is no data published for the phylogeographic approach. Based on archaeology, cultivated flax reached the Alpine region as part of the agricultural assemblage introduced by early farmers who emigrated from Asia Minor (Helbaek 1959). The earliest record of linseed in Europe was found in Greece and is dated back to around 6000 years BCE (Valamoti 2011). Later archaeobotanical evidence was discovered in Late Neolithic sites of Slovenia (Tolar *et al.* 2011). Flax seeds were also found in Central European Lake Dwellings where between 4000 and 2500 years BCE decrease in seed size was observed (Herbig & Maier 2011). Moving further north, the earliest archaeological findings of flax in Denmark are dated back to 1800 BCE and then it remains recorded continuously until now (Karg 2011). Spatio-temporal distribution of flax artefacts is congruent with the general understanding of the early agriculture spread described in paragraphs above.

The remains of the agrarian and processing technology associated with flax have shorter history than flax cultivation itself and hence its utility in researching its spread is limited. The most comprehensive study on both flax and associated technology remains in Central Europe shows that flax was grown and used for linen production in the Lake Dwellings of Upper Swabia and Lake Constance throughout the Late Neolithic (Maier & Schlichtherle 2011). In this period of time there were many innovations observed in the archaeological layers including new harvesting tools, retting pits and spindle whorls that were used for better fibre processing. The technology of flax textile production used by the early Europeans is well understood

thanks to a combination of archaeological investigation and electron microscopy of linen remains of the time. Fibres were retted but not combed and thread was plied using two yarns (Leuzinger & Rast-Eicher 2011). In summary, there is strong evidence for specialization in textile production and increased importance of flax in Late Neolithic Central Europe. Later, these technologies have spread to Europe and Asia.

1.2.4 Flax varieties and their economic importance

Cultivated flax is grown for fibre from its stems and for oil extracted from seeds. Specialized varieties are bred for either of the two products. Hence, the principal use of varieties along with morphological traits are the criteria used in grouping flax varieties into the four major convarieties described by Kulpa & Danert (Kulpa & Danert 1962). These convarieties are: *crepitans*, *mediterraneum, elongatum*, *usitatissimum* and they correspond to the dehiscent, large-seeded (associated with oil production), long (associated with fibre production) and intermediate varieties respectively. The dehiscent flax is characterised by spontaneously opening capsules, which is believed to be a primitive trait. The large-seeded convariety has side floral branches formed along more than half of the stem height and its seed weight exceeds 9 g per 1000 seeds. Linseed contains on average 38.3% oil. Diederichsen & Raney (2006) showed that the seed size is positively correlated with oil concentration rendering *mediterraneum* varieties the best for oil production. Also, varieties that have yellow seeds are characterized by higher oil content on average. The long varieties are typical fibre varieties with fewer seeds, long stems and with side floral branches that are formed in less than 1/4 of the overall plant height. On average, dried flax stems contain 21.4% of fibre and this value is correlated with a height of plant and its branching pattern (Diederichsen & Ulrich 2009) that makes the convariety *elongatum* the most efficient for fibre production. *Usitatissimum* varieties have the intermediate traits and could be grown for fibre and/or oil. However, improvement of oil yield usually results with negative effect on fibre yield and *vice versa*. Hence, the specialized varieties are preferred.

Cultivated flax is a plant of great economic importance both as a food crop and a source of raw materials for industrial processes. The oil produced from flax seeds is rich in omega-3 fatty acids (Vereshch & Novitska 1965) and is used as a nutritional

supplement for animals and humans. Linseed oil is also reported to act as anti-inflammatory compound (Zhan *et al.* 2009). Its most common application is in production of wood impregnators and varnishes. Historically, for oil production, seeds were removed from the plant by threshing. Next, oil was produced by pressing seeds and sieving to remove seed coat and other hard material. In the present day this process is improved by the chemical solvent extraction. Over the last 50 years production of flax oil decreased by almost half (Table **1.1**).

Table 1.1: Flax seed and fibre production (in tonnes) over the last 50 years in the world based on FAOstat calculations.

	2011	2001	1991	1981	1971	1961
Seed	1602047	1875277	2691299	2246614	3286032	3014416
Fibre	315084	622645	731637	609564	776566	696579

Fibre from flax is used to produce the fabric linen, which has very high quality textile properties of strength, water absorption and texture. Flax fibre is also used to develop natural reinforced composites (Kalia *et al.* 2009). Transgenic flax is also used as an anti-oxidising material for bandages that improve treatment of ulcers (Skorkowska-Telichowska *et al.* 2010). Historically fibres were extracted from flax stems in a process of retting. Stalks were left in waterlogged retting pits to allow for organic decomposition of soft tissues. Next, the material was dried and impurities removed in a process called scotching. Finally shorter fibres and straw were removed with use of heckling combs and the remaining tow was spun into yarns. Nowadays, this process is mechanised and much more efficient. In the past ten years flax fibre production has decreased significantly (Table **1.1**), however, there are multiple innovative studies that could potentially increase importance of this crop on textile markets.

Despite a decreasing importance of flax in cloth production, there is demand for its intensified cultivation. Currently, the major goals of flax breeding programmes are to improve its crop quantitative and qualitative traits and pathogen resistance, especially to the rust (Catanzariti *et al.* 2009), the bacterial scorch (Cariou *et al.* 2003) and other

fungal diseases (de Wit *et al.* 2009). However, scant genetic resources of flax landraces and genetic uniformity of most of the cultivars remain a major problem. The intermediate varieties constitute the great majority of the flax germplasm in Canadian Genetic Resources and are predominant in the German National Gene Bank, while varieties belonging to any of the two specialized groups are relatively rare (Diederichsen 2009).

According to the recent phenotypic and molecular analysis, genetic diversity of fibre flax varieties is generally low. This observation was initially made during the assessments of germplasm diversity when it was found that the flax cultivars do not match the criterion of distinctiveness (Everaert *et al.* 2001). Distinctiveness test together with uniformity and stability tests are necessary in variety identification. Failing the distinctiveness test means that there are no traits that allow distinguishing one variety from the others. AFLP analysis confirmed a low diversity of the flax accessions in general and later, extended studies showed that this problem is especially relevant to fibre flax varieties (Vromans 2006). A comparative study revealed that phenotypic diversity in the North American flax cultivars decreased during the 20th century (Diederichsen 2009). The possible cause of the low diversity might be attributed to a neglect of breeding programmes and efforts to conserve biodiversity of flax varieties for nearly two centuries. This was a result of increased popularity of cotton that replaced linen as the most important textile plant in both North America and Europe (Solar 2012). This period of time could be perceived as a population ‘bottleneck’ for flax cultivars. Poor conservation of diversity hindered breeding programmes for this economically important crop plant.

1.3 RATIONALE FOR PROJECT HYPOTHESIS

1.3.1 Specialized flax varieties

Of all the founder crops domesticated in the Near East, flax is unique in its dual use of fibres for textile production and seeds as a source of oil. However, the original purpose of flax cultivation and the time of emergence of varieties are unknown. It was stated in Section ***1.2.4*** that modern flax varieties are classified into the four convarieties: fibre, oil, intermediate and dehiscent flax. In studies on the flax evolution history additionally a fifth important group is characterized – the winter varieties. The winter varieties have very similar phenology to pale flax; they grow in winter to flower in early spring and as such fit the climatic conditions in the Near East rather than in Europe. Both winter flax and convariety *crepitans* are thought to be products of early domestication processes. There is evidence that plants classified as convariety *crepitans* are more ancient than others (Fu 2011, 2012; Uysal *et al.* 2010) and that winter varieties are much more closely related to pale flax than to modern cultivars (Fu 2012). In turn, convariety *usitatissimum* with its non-shattering capsules was the first type of flax that was completely domesticated and possibly grown for both seeds and stems. Finally, convariety *mediterraneum* and convariety *elongatum* are the specialized forms that probably arose later. It is remarkable that single species was used for the two purposes, especially given that specialization in fibre production reduces high oil yield and *vice versa*.

Origins of convariety *mediterraneum* and *elongatum* are unknown but they are speculated to have arisen outside the original domestication area. It is very unlikely that any specialized forms were cultivated in the Neolithic Near East, however, some molecular analyses suggest that flax might have been domesticated for its oil rather than for fibre (Allaby *et al.* 2005; Fu & Allaby 2010). Flax seeds appeared first in ancient human settlements from 9000 years BCE (Hillman 1975) and since then are recorded continuously in the archaeological sites of the Old World. In the Near East the earliest discovered putative textile remains were dated to 7000 BCE but linen itself is archaeologically documented in Egypt only as late as 2400 BCE (Vul'f & Elladi 1940) and historically is mentioned in the Old testament. The archaeological evidence suggests that cultivation of specialized, fibre varieties of flax started in Neolithic Alpine Lake Dwellings in Central Europe. Cultivated flax reached Alpine region as part of the agricultural assemblage introduced by the early farmers who

emigrated from Asia Minor (Helbaek 1959). Two interesting phenomena occurred in Alpine archaeological records of this period:
1) a shift in the size of flax seeds from large in the lower strata (associated with oil production) to smaller in the top ones (associated with fibre production) (Herbig & Maier 2011) and
2) abundance of the agrarian and technical innovations that enhanced textile production (Maier & Schlichtherle 2011).
These findings suggest that the European Neolithic communities specialized in growing flax for its fibre, and used a distinct variety (similar in seed size to convariety *elongatum*) for this purpose.

1.3.2 Adaptation to the northerly latitudes and emergence of fibre flax varieties

The shift in seed size and consequent emergence of fibre flax forms in Central Europe might be associated with an adaptation to shorter vegetative periods and more extreme winters, which are typical in the northern latitudes. A correlation between the latitude and seed size was shown for example in *Arabidopsis thaliana* (Li et al. 1998). Better understanding of the mechanisms of plant adaptation to the European climate might shed some light on its impact on plant morphology.

Change in latitude comes with major change in climatic conditions that can force crop plants to adapt via modifications in their life cycle – often flowering time and growth determinacy. In balsam poplar an effort to resequence 27 flowering genes in samples from wide range of latitudes was made (Keller *et al.* 2011a). Despite a very weak signal of selection at the whole-species level, strong local selection between northern and southern subpopulations was discovered mainly in four flowering loci. Similarly, in European Aspen seven out of 25 investigated flowering genes were under positive selection (Hall *et al.* 2011) while in *A. thaliana* three genes were found to be major contributors towards the adaptation to different environments (Flowers *et al.* 2009). This evidence shows that indeed flowering time genes played an important role in latitudinal adaptation in many plant species.

In crop plants, adaptation to northern latitudes is linked to change of the two phenological traits: reduction of vernalization requirements and photoperiod insensitivity (Fuller & Allaby 2009). Vernalization is the typical control mechanism in the Near Eastern winter crops. It allows plants to maximize vegetative growth over

cool and wet winter conditions, which are characteristic for the Near East. Vernalized plants flower early in the spring to avoid drought conditions often associated with summer (Cockram *et al.* 2007). If the cold period is too short, then winter variety plants will not flower. Additionally, crops often respond to the daylength in order to control the start of flowering. Long day (LD) plants will promote flowering only if the threshold of their sunlight perception period is met, which usually translates into more than 12 hours. Facultative LD plants will suffer from delayed flowering if this condition is not met, while obligatory LD plants will not flower at all. This is the second, very important mechanism in timing of flowering and its role is again to delay flowering until spring to optimize growth in the Near Eastern climate conditions.

Both mechanisms of flowering control are dependent on environmental cues. Thus, winter crops, which are adapted to the conditions in the Near East, would struggle to optimize flowering time in the North, where the climate is very different. Limitation in the geographic distribution of winter varieties is well illustrated with the distribution of wheat in the USA. Winter varieties of bread wheat are grown in the southern states over the winter, while spring varieties and durum wheat are grown in the northern states over the summer (Leff et al. 2004). Furthermore, winter barley varieties are cultivated in most of Europe, while spring varieties are preferred in North-Eastern Europe (Poland, Belarus, Baltic countries and Russia). The northern climatic conditions might be difficult for winter crops, especially in presence of harsh winters.

Crop plants that were introduced to the northerly latitudes might have adapted through change in the growth habit. It is likely that there was a selective pressure towards crops sown in the early spring and, in consequence, towards reduction of vernalization requirement. Crops adapted in this way could become spring varieties. Spring varieties could flower later relative to the calendar year allowing for minimum required vegetative growth. The same daylength will occur earlier in the calendar year with increasing latitude and hence due to promoting signal from photoperiod-dependent pathway winter varieties in the north will flower earlier. Hypothetically, crop plants that adapted to the Northern climate through reduction of vernalization requirement are likely to have evolved insensitivity to daylength or flowering inhibition mechanism that would extend the vegetative phase.

Alternatively, these plants could switch into indeterminate growth habit and flower continuously.

Most of the Near Eastern founder crops adapted successfully to the European climate. For example the mechanism of flowering in cereals and its role in adaptation is well understood. Wheat and barley have orthologous genes controlling their photoperiodicity. Five genes have been associated with the modifications in the cereals' seasonality: *VRN1*, *VRN2*, *VRN3* and *PPD1* reviewed in Fuller & Allaby (2009), and recently the fifth - *HvCEN* added by Comadran *et al.* (2012). A single mutation in *PPD1* can result in the haplotype *H1* that is associated with the adaptation to northern latitudes and probably arose in the wild barley populations of the Iranian highlands (Jones *et al.* 2008). In hulled emmer wheat *PPD-A1*(I) haplotype that differs by indels from *PPD-A1*(II) is associated with the landraces cultivated north from the Mediterranean region (Takenaka & Kawahara 2012). In bread wheat flowering time could be altered by only manipulating the number of copies of *PPD-B1* and *VRN-A1* (Diaz *et al.* 2012). It has been concluded that mutations in *PPD* genes led to the adaptation to the northerly latitudes usually through a change to constitutive activation of photoperiod pathway (Cockram *et al.* 2007). Photoperiod adaptation was also investigated in pea, the only legume plant that was successfully introduced to Europe. Major loci controlling the differences in photoperiod response were identified with the most of the impact attributed to gene *HR* (Weller *et al.* 2012).

The second element of adaptation – the loss of vernalization response is well understood in the cereals and is attributed to mutations in *VRN* genes. It has been shown that *VRN*-A1 haplotype in einkorn wheat and *VRN*-A1, *VRN*-B1, *VRN*-D1 in bread wheat are associated with spring varieties (Iqbal *et al.* 2007). In barley exclusively, *HvCEN* haplotypes allow distinction of winter and spring varieties (Comadran *et al.* 2012).

The strength of selection pressure for these changes depends on how far north the early farmers migrated before adaptation of crop plants. According to meta-data analyses the selection pressure for change in flowering time occurs at higher rates with higher latitude (Munguia-Rosas *et al.* 2011). Diederichsen & Hammer (1995) suggest that the replacement of winter varieties occurred in Central Europe

somewhere along the Danube River and hence the selection pressure might have been very strong already. Lentils and chickpeas are crops domesticated in the Near East that did not adapt during the early farmers' movements and thus were not grown historically in Europe (Concolly *et al.* 2008). These two pulses are described as intolerant to reduction of growing season and increased autumn moisture (Smartt 1990), which might explain difficulties in adapting to the northerly latitudes under the strong selection regime. Other crop plants were more tolerant to reduction of growing season and hence survived the selection. It was not an easy task however, as the delays in agricultural spread could be a reflection of the time taken for crops to adapt to new climatic conditions (Jones *et al.* 2007).

There is little known about adaptation of cultivated flax to the northerly latitudes. Flax is one of the latest-maturing traditional crops and hence the climate in the northern zone increases the risk of frosts before seed maturation (Dribnenki 2010). For this reason plants might have been forced to produce seeds earlier in the season by the cost of their reduced size, which in the case of cultivated flax led to lower oil yield. This would be congruent with archaeological findings on decreasing flax seed size in Alpine Neolithic strata (Herbig & Maier 2011). Consequently, it may be the case that the intermediate flax varieties, which arrived from the Near East adapted to the Northern latitudes, but compromised their oil-producing properties. Furthermore, northern-grown Canadian flax is characterized by indeterminate growth habit. It has been shown for soybean that indeterminacy is associated with the plants cultivated in the northern parts of China while southern plants are characterized by determinate growth habit (Tian *et al.* 2010). In flax, indeterminacy results in late maturity and green stems after bolls maturation (Dribnenki 2010). Harvesting green rather than dried flax stalks might have had major impact on fibre flexibility and in consequence improved textile quality. In flax, the Near Eastern varieties adapted to northerly latitudes probably by the cost of reduced seed size, however, might have serendipitously improved fibre production that could have driven a shift of use of this crop. Both indeterminacy and flowering time are controlled by the same set of flowering genes and hence they should be targeted as potentially adaptive loci in flax.

1.3.3 Potential role of pale flax in adaptation

Domestication of crop plants in the Near East could be understood as a process in which early farmers wielded the wild plants and shaped them into more useful cultivated equivalents all in one go, in the narrow brackets of time and space. Arguments in favour of this scenario (rapid transition) are often debated in literature. Hillman and Davies (1990) suggested that the domestication process was very rapid based on their mathematical modelling and field trials study. Furthermore, the monophyletic character of crop plants populations speaks in favour of centralised and quick domestication (Zohary 1999), however, it has been shown that phylogenetic analyses are prone to mistakes in delimiting single and multiple origins in crop species (Allaby *et al.* 2008). An interesting observation was made in favour of rapid transition: an effective population size of less than 1,500 individuals is enough to explain the diversity seen in modern crops such as rice and maize (Eyre-Walker *et al.* 1998; Zhu *et al.* 2007). This does not mean however, that it was in fact that small. The debate over the pace of domestication process is not yet fully resolved, but evidence is accumulating for the protracted transition paradigm.

The theoretical model of domestication has changed with time. We now are beginning to see it as a long and complex process that smoothly developed into modern breeding practices rather than as a rapid transition. In this scenario domestication was preceded by pre-domestication, which started much earlier than 10000 years BCE and involved growing gathered plants next to human settlements. Syndrome traits were fixed into wild plants during domestication. This stage was followed by the spread and adaptation of crop plants to local conditions, dated back to 8000 - 4000 years BCE. After that crops were changed to increase their nutritional properties, an effort that continues in the modern times (Brown *et al.* 2009). There are multiple arguments for protracted paradigm: 1) the latest computer simulations (Allaby *et al*. 2014), 2) slow rates of selection under domestication (comparable to these observed for wild species) based on a large dataset consisting of the eleven species in the 60 archaeological sites (Purugganan & Fuller 2009; Purugganan & Fuller 2011) and 3) multiple origins of some of the crop species based on molecular data (Brown 2009).

Several crop species have been shown to have multiple origins based on independent acquisition of the domestication syndrome or climate adaptive traits. In barley, for example, indehiscence is linked to tough rachis mutation; since there are two separate mutants of this domestication syndrome trait that provided evidence for multiple origin of this crop plant (Azhaguvel & Komatsuda 2007; Takahashi 1972). Secondary changes after domestication are often driven by adaptation to various local conditions. In barley, it has been shown that a two-row variety emerged independently in Egypt as a result of an adaptation to arid environment (Palmer *et al.* 2009). Furthermore, varieties that are associated with the North inherited their adaptive trait from the wild populations inhabiting Iranian Highlands (Jones *et al.* 2008). Multiple origins of barley domestication where thoroughly investigated because this cereal was used as a model plant for studying domestication. There is however, emerging genetic evidence that other Near-Eastern crop plants such as einkorn (Kilian *et al.* 2007) and emmer wheat (Ozkan *et al.* 2005; Ozkan *et al.* 2011) also had multiple origins.

Independent acquisition of the climate adaptive traits, which supports a protracted paradigm, might be caused by secondary contact between domesticate and its wild relative. Of all the crops in the 'Neolithic package', flax is unusual in that the biogeographic range of its wild relative stretches well beyond the Near East into the northerly latitudes. It inhabits Croatia, Slovenia, Northern Italy, the most of France and South-western shores of the United Kingdom and Ireland (Vul'f & Elladi 1940). It is highly probable that these populations evolved an adaptation to the northerly latitudes. Uysal and collaborators (2010) argue that genetic diversity observed in pale flax is suggestive of strong local adaptation. During the spread of cultivated flax to Central and Northern Europe these adaptations could have been passed from pale flax populations through hybridization, since as it has been shown – at low rates flax can cross-pollinate (Gurbuz 1999) and both species are interfertile (Gill & Yermanos 1967; Muravenko *et al.* 2003). If this is a case, cultivated flax could be used as a model plant for protracted domestication paradigm.

1.3.4 The research hypothesis

Cultivated flax with its dual use and wild relative that is dispersed in broad biogeographic range is an excellent model plant for evaluating the protracted domestication paradigm. It is possible that wild European populations of flax contributed significantly towards the adaptation of cultivated flax to the Northerly latitudes and the emergence of the specialized fibre varieties. In this thesis I am posing the hypothesis that:

Northern-adapted, fibre varieties of flax have had significant input from northern latitude wild flax populations in support of the scenario that the evolution of fibre use is associated with northern gene flow. The alternative scenario is that:

Northern-adapted, fibre varieties of flax are a subset of the diversity of domesticated oil producing flax and are the product of selective breeding within that gene pool in support of the notion that the evolution of fibre use was independent to wild gene flow in cultivated flax.

To test these hypotheses, genetic markers associated with flowering time have to be developed for investigation in flax. Furthermore, additional neutral markers have to be developed to estimate the gene flow rates between cultivated and pale flax.

1.4 MOLECULAR MECHANISMS OF FLORAL INITIATION

1.4.1 Molecular mechanisms of floral initiation in Arabidopsis thaliana

In wild and cultivated annual plants flowering time is an important life-history trait. It coordinates the transition from vegetative growth to flowering with environmental stimuli such as photoperiod, temperature and biotic factors. Using this system, plants are capable of deciding on the optimal time for generative phase, ensuring that there are enough resources to be allocated in progeny and that atmospheric conditions are favourable for its survival. Furthermore, this system allows plants to either flower continuously through their lifespan (indeterminate growth) or focus on producing progeny in determined time (determinate growth). Due to its wide geographic distribution, well-studied ecology and genetics, *A. thaliana* was accepted as model plant for studies on the flowering time control mechanism on molecular level.

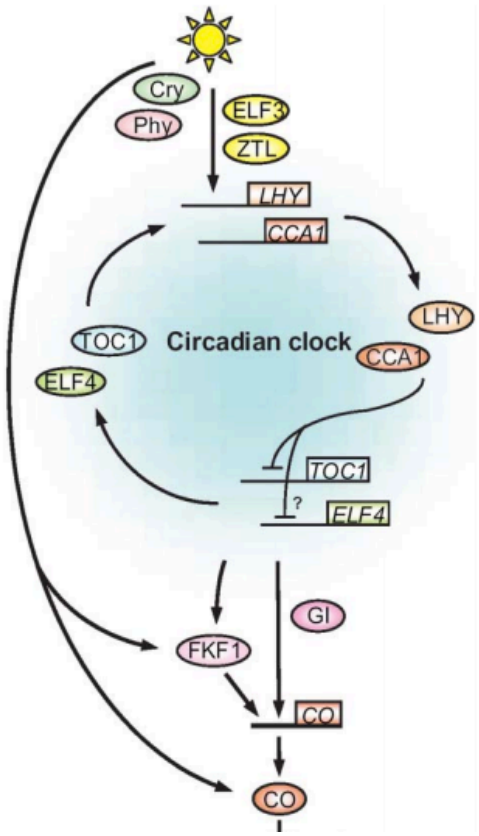


Figure 1.1: Model of the circadian system in *Arabidopsis thaliana*. Photoperiod affects the circadian clock and directly up-regulates *CO* expression (Hayama & Coupland 2004)**.**

Floral meristem identity genes fulfil the key function in the flowering transition process of A. thaliana. Two loci, *APETALA1* (*AP1*) and *LEAFY* (*LFY*), are the most important genes for floral initiation. They both belong to *MADS*-box gene family and serve as transcription factors that switch the flower development machinery on (Mandel *et al.* 1992; Weigel *et al.* 1992). Both *AP1* and *LFY* are directly controlled by group of genes called flowering signal integrators. This group consists of floral activators, such as *FLOWERING LOCUS T* (*FT*), *TWIN SISTER OF FT* (*TSF*), *MOTHER OF FT AND TFL* (*MFT*), *FLOWERING LOCUS D* (*FD*) and inhibitors – *TERMINAL FLOWERING LOCUS 1* (*TFL1*), *BROTHER OF FT AND TFL1* (*BFT*)

and *Arabidopsis thaliana CENTRORADIALIS* (*ATC*). Upstream from these genes are the four regulatory pathways: gibberellin-, vernalization-, light-dependent and autonomous (Putterill 2001). In this thesis adaptation to northerly climatic condition is investigated and so only the two pathways that are directly controlled by environmental factors are discussed: the light- and vernalization-dependent pathways.

The light-dependent pathway is also referred to as photoperiod-dependent because it is the perception of daylength rather than light itself that regulates flowering promotion. Plants can promote flowering in short days (SD), long days (LD) or be neutral, which means that they are not responsive to daylength. *A. thaliana* is classified as LD plant and hence longer days (usually more than 12 hours) will enhance its flowering. Daylength perception is coordinated with the circadian clock – an internal timing mechanism (Figure **2.1**). These two mechanisms were first linked in the study of *A. thaliana LATE ELONGATED HYPOCOTYL* (*LHY*) and *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) genes (Schaffer *et al.* 1998; Wang & Tobin 1998). The two genes together with *TIMING OF CAB 1* (*TOC1*) are considered as the oscillator loop that keeps track of passing time (Alabadi *et al.* 2001). Downstream, the clock is responsible for timing expression of transcription factor *GIGANTEA* (*GI*), which in turn up-regulates key gene in photoperiod-dependent floral initiation – *CONSTANS* (*CO*) (Putterill *et al.* 1995). Mutation in *GI* confers photoperiod insensitivity (Park *et al.* 1999). There is evidence that the *GI* mutant effect is caused by an impact on the circadian rhythm and that it directly up-regulates *CO* gene expression (Mizoguchi *et al.* 2005). In wild type only long days activate *CO* expression, however, its protein is stable only in presence of light. It means that active proteins are present in plant cells only for short period of time between late day expression and the sunset. When stable, *CO* up-regulates *FT* and *SUPPRESOR OF OVEREXPRESSION OF CONSTANS* (*SOC1*), these integrator genes promote flowering transition (Suarez-Lopez *et al.* 2001). Altogether the circadian clock constitutes a very precise mechanism for timing the flowering, yet it is very flexible in adaptation.

The second, environmentally dependent genetic pathway is regulated by temperature. *A. thaliana* ecotypes collected from European latitudes often flower more rapidly when their seedlings are exposed to a prolonged cold-treatment. This process is

known as vernalization. Two major loci determine flowering time in vernalization-dependent plants: *FRIGIDA* (*FRI*) located on chromosome 4 (Clarke & Dean 1994) and *FLOWERING LOCUS C* (*FLC*) on chromosome 5 (Koornneef *et al.* 1994). Dominant alleles of these genes serve as the mechanism of flowering control. Both will delay flowering unless the plant has been vernalized. The only known function of *FRI* is up-regulating *FLC*. By contrast to *FRI*, another regulatory gene, *VERNALIZATION 2* (*VRN2*) inhibits *FLC* expression (Sheldon *et al.* 1999). Downstream, *FLC* down-regulates integrator gene *SOC1* (Samach *et al.* 2000) and directly inhibits *FT* expression delaying the flowering transition. The model of vernalization-dependent pathway assumes that cold treatment causes demethylation of *FLC* reducing its expression and in consequence its inhibiting power (Sheldon *et al.* 2000). In effect, cold treatment is promoting flowering in vernalization-dependent plants.

The genes described above form only the core of flowering gene network. Over time, the model of genetic interactions among flowering genes has become more and more complicated. Ehrenreich and collaborators (2009) present the interactions within the complex flowering network based on compilation of previously published genetic studies. Additionally they investigated the impact of 48 genes on flowering time through association mapping. Their results suggest that 4 to 14% of known flowering genes have frequent polymorphism that contributes towards natural variation of flowering time in *A. thaliana*.

1.4.2 Flowering time genes in poplar and flax

The core of flowering gene network is conserved in many plant species. For example, the *CO* and *FT* regulation system is very common within eudicots and occurs even in monocots such as rice and barley (Fuller & Allaby 2009). The closest related model species to flax, in which photoperiod-dependent pathway was extensively studied, are poplars (*Populus* sp.). Poplars are woody perennials that flower in long-days. In *P. balsamifera* homologs of 27 flowering-time genes were resequenced, including well-studied loci such as *LFY*, *CO* and *TFL* (Keller et al. 2011a). Furthermore, in *P. deltoides CO* homologs were found to be accommodated by two loci which are named *CO-LIKE1* and *CO-LIKE2*, both likely to be involved in promotion of flowering (Yuceer *et al.* 2002). Next, in *P. nigra* five homologs of

FT have been discovered; in total these species have nine members of the *FT*/*TFL1* gene family (Igasaki *et al.* 2008). In *P. tremula*, *PtFT1* protein expression is activated by *PtCO2* when daylength reaches a certain threshold level (Bohlenius *et al.* 2006); in this way it confers the same function as their homologs in *A. thaliana*. At the same time, both *FT1* and *FT2* expression is dependent on the ambient temperature (Hsu *et al.* 2011). Interestingly, no functional or structural paralogs of *FLC* have been found in poplars (Leseberg *et al.* 2006) and hence it was hypothesized that other genes assumed *FLC*'s role in temperature-dependent control of flowering. Both temperature and photoperiod-dependent pathways of flowering control were found important in latitudinal adaptations of poplar (Bohlenius *et al.* 2006; Keller *et al.* 2011b). In summary, function of many *A. thaliana* flowering genes is conserved within poplar.

In flax, the flowering-time control mechanism is largely unknown. Very little is known about induction of flax floral development on molecular level. Two putatively homologous genes were found in flax genomic DNA: *LFY* and *TFL1*, while only the former was found in the cDNA (De Decker 2007). Expression of other floral gene homologs was confirmed in flax, namely: *SOC1*, *CO*, *ADG1*, *GAI* and *AP1* (House 2010). The other important genes such as *LHY*, *CCA1*, *TOC1*, *GI*, *FT*, *FRI*, *VRN2* and *FLC* were either not investigated or not found in flax genome. Flax has been reported as LD plant that often is vernalization-dependent (Domantovich *et al.* 2012; House 2010; Thomas *et al.* 2006). Hence, its molecular mechanism of flowering could be compared to *A. thaliana* model. If flowering genes are conserved between *A. thaliana*, flax and poplars we should expect the presence of *CO*/*FT* system in photoperiod-dependent pathway in flax. On the other hand, since flax is closely related to poplars we should expect the absence of *FLC* in temperature-dependent pathway.

In North America, flax varieties that are associated with the south are responsive to vernalization and photoperiod while the northern counterparts are insensitive to vernalization. Furthermore, on average northern varieties flower earlier (Darapuneni et al. 2014). Fieldes and Amyot (1999) constructed a model of early flowering in flax, which assume that at least three loci are involved in the timing of the flowering. These loci are responsible for: 1) flowering promotion, 2) plant height and 3) control of the two aforementioned. This bares remarkable similarity to *CO* and *FT* system,

where *CO* controls *FT*, which in turn promotes flowering and regulates seasonal growth in related poplar (Bohlenius et al. 2006). The last locus could be *TFL1* with its impact on flowering time and plant height. Furthermore, there are reports that methylation might be playing an important role in controlling flowering time in flax (Brown et al. 2008; De Decker 2007; Fieldes & Harvey 2004). In summary, there is an evidence for latitudinal gradient of flowering behaviour in flax. There are theoretical premises that phenotypic diversity along this gradient might be controlled by loci similar to *CO*, *FT* and *TFL1*. To the best of the author's knowledge these studies represent our knowledge of flowering control mechanism in flax to date.

1.4.3 The PEBP family genes

FT and *TFL1* genes belong to a multi-gene family including genes that encode proteins with structural similarities to mammalian phosphatidylethanolamine-binding protein (PEBP) (Schoentgen *et al.* 1987). In *A. thaliana* there are six genes that encode PEBP-like proteins and provide function as either floral inhibitors (*BFT*, *TFL1* and *ATC*) or promoters (*FT* and *TSF*), and each having a characteristic expression pattern. For example, *FT* is expressed only in the phloem of *A. thaliana*, and is up-regulated by *CO* during long days by the evening (Kobayashi *et al.* 1999). *TSF* is also up-regulated by *CO* and expressed in the phloem, but differs from *FT* in the expression pattern during the seedling stage (Yamaguchi *et al.* 2005). The third gene *MFT*, also promotes flowering when overexpressed, however, its impact is negligible under natural conditions (Yoo *et al.* 2004). In contrast, *TFL1*'s role is to delay flowering (Shannon & Meekswagner 1991). It is expressed in apical meristem throughout vegetative stage (Bradley *et al.* 1997). Interestingly, some studies report that *TFL1* is also regulated by *CO (Simon et al. 1996)*. *A. thaliana* has two paralogs of *TFL1* with similar activity named *BFT* and *ATC*. *BFT* is characterized by weaker impact on flowering time and counter-intuitively shows an expression pattern similar to *FT* that peaks in the evening (Yoo *et al.* 2010). *ATC* shows an expression pattern similar to a homolog in *Anthirrum* sp. called *CENTRORADIALIS* (*CEN*). Both genes are expressed in the inflorescence apex after floral induction (Bradley *et al.* 1996; Mimida *et al.* 2001), however unlike other genes, are up-regulated by short days (Huang *et al.* 2012).

1.4.4 The adaptive potential of PEBP family genes

The adaptive potential of flowering genes from the PEBP family is remarkable. They often acquire new functions through gene duplications and mutations. Additionally, the same PEBP genes could be regulated by different cues in different species. In poplar, there are multiple *FT* paralogs (Igasaki *et al.* 2008), at least two of them with conserved function, however, one (*FT1*) responds to low while second (*FT2*) to high temperatures (Hsu *et al.* 2011). Three *FT* homologs with conserved function were found in *Medicago tranculata*, one of which (*MtFT1a*) was responsive to cold temperatures while all three were responsive to long days (Laurie *et al.* 2011). In beetroot (*Beta vulgaris*) *FT* was duplicated and one of the derived genes (*BvFT1*) mutated and changed function from floral promoter to inhibitor. Interestingly, both *BvFT1*and *BvFT2* are still sensitive to vernalization and daylength (Pin *et al.* 2010). Similarly, in sunflower (*Helianthus annuus*) frame-shift mutation in one of the *FT* paralogs (*HaFT1*) led to a delay of flowering (Blackman *et al.* 2010). An up to date list of identified *FT* homologs in different plant species was presented by Matsoukas and collaborators (2012).

A variety of studies report that *FT*/*TFL* system offers great plasticity in adaptation to various climates, plasticity that might have been employed during domestication and early stages of spread and adaptation of crops to local environments. The *HaFT1* gene in sunflower experienced a selective sweep during domestication that resulted with change in this crop's phenology (Blackman *et al.* 2010). In domesticated tomato *FT* homolog *SP5G* is inactive due to mutation causing early stop codon (Tomato Genome Consortium 2012). Similarly, *SP6A* is expressed only in wild tomato (Carmel-Goren *et al.* 2003). In soybeans determinacy is an important trait associated with domestication and it has been shown that mutations in *GmTFL1* locus are potentially adaptive to northern latitudes (Tian *et al.* 2010). Furthermore, a homolog of *CEN* in barley was shown to be associated with summer varieties and an important factor in adaptation to European environment (Comadran *et al.* 2012). These examples show the importance of *FT*/*TFL* genes in domestication and local adaptations of crop plants.

1.4.5 PEBP genes impact on growth determinacy and plant architecture

Plants with determinate growth habit will stop growing once a flower has completely formed. Further growth is possible only by initiating axillary meristems (sympodial growth). By contrast, indeterminate plants continue their apical growth despite forming flowers, which in this case are formed in axillary buds (monopodial growth). These plants continue flowering until frosts or other environmental factors terminate them. Crop species, such as tomato, can exhibit both determinate and indeterminate growth habit.

One of the most important consequences of the different levels of *FT*/*TFL* expressions and different flowering times are their impact on plant floral architecture and switch from determinate to indeterminate growth (Alvarez *et al.* 1992). In tomato, it has been shown that *FT*/*TFL* ratio will have an impact on switch between sympodial/monopodial type of growth and inflorescence promotion (Lifschitz *et al.* 2006; Shalit *et al.* 2009). Similarly, *TFL1* homolog in peppers have an impact on plant architecture (Elitzur et al. 2009). In their model for *A. thaliana*, Prusinkiewicz and collaborators (2007) explained how *FT*/*TFL* ratio could lead to development of different inflorescence types and explored their adaptive properties in various environmental conditions. The indeterminacy could be an important trait leading to an adaptation to the northerly latitudes, where growing season is more variable. For example, it has been reported that a *GmTFL1* mutation causing indeterminate growth in soybeans is associated with plants growing in the northerly latitudes and therefore was important in adaptation to local conditions (Tian *et al.* 2010). Furthermore, in perennial aspen *FT* is responsible for growth cessation over winter and is critical in adaptation to northerly latitudes (Bohlenius *et al.* 2006). The conclusion drawn from these examples is that determinate plants are favoured in the south where the duration of the growing season is not a limiting factor. However, in the north where shorter growing period and more variable conditions occur, indeterminate plants fare better. This growth habit allows them to establish reproductive primordia continuously throughout the season in attempt to maximize production of progeny before winter. Flax in Canada grows indeterminately and its stems often remain green and flexible even after seed maturation (Dribnenki 2010), which potentially improves fibre production. It might be the case that the change in flax's growth habit in the northerly latitudes was a result of mutations in PEBP genes.

There is evidence for many crop plants that PEBP genes have an impact on other economically important traits through controlling determinacy, plant architecture and in consequence - resource allocation. One of the *TFL1* homologs in rose – *RTFL1c* is probably responsible for flowering termination; when not expressed it causes a desired, recurrent flowering phenotype (Wang *et al.* 2012a). PEBP genes moreover, have an impact on fruit yield and other agronomical qualities. A specific combination of *SFT* alleles in tomato will cause heterosis effect, which leads to approximately 60% increase of yield (Krieger *et al.* 2010). Furthermore, it has been suggested that manipulation of *FT* expression in cotton could improve its architecture for agronomic needs (McGarry & Ayre 2012a). In summary, there is mounting evidence for the importance of PEBP family genes in the process of local adaptation but also for their role in changing agronomic traits in crop plants. These genes might have played an important role in flax's adaptation and yield improvement at the same time. It has been noted that one of the three loci responsible for early flowering in flax also controls plant height (Fieldes & Amyot 1999) – a major yield trait in fibre varieties.

1.5 NON-NEUTRAL AND NEUTRAL MARKERS, GENOTYPING METHODS AND ASSOCIATED STATISTICAL ANALYSES

1.5.1 Conventional molecular markers

Molecular markers are fragments of protein or DNA molecules with naturally occurring polymorphism. This polymorphism is used in determining the difference in genetics of individuals, populations, species or higher taxonomical ranks. Molecular markers are commonly used in characterization of genetic variability, genome fingerprinting, genome mapping, gene localization, analysis of genome evolution, population genetics, taxonomy, plant breeding and diagnostics. There is a set of properties that constitute a good molecular marker. It should be co-dominant (which means that all forms of the marker will be discriminated in polyploid organisms), frequent in the genome, easy to assay and reproducible.

Conventional molecular markers could be divided into categories based on the enzymatic reactions that are used to analyse them. First, there are non-PCR markers such as restriction length polymorphism (RFLP). In RFLP, a restriction reaction with a specific enzyme is used to digest genomic DNA and the fragments are visualized using gel electrophoresis. Different fragmentation patterns reflect mutations in restriction sites. Secondly, there are PCR-based markers. In amplified fragment length polymorphism (AFLP) method, restricted fragments of DNA are selectively amplified in a PCR reaction (Vos *et al*. 1995). In randomly amplified polymorphic DNA (RAPD), fragments of DNA are amplified using short, random primers. Different fragmentation patterns on a gel reflect mutations in priming sites. RFLP, AFLP and RAPD are not co-dominant markers. In simple sequence repeat (SRR) method, earlier characterised microsatellite loci are amplified (Queller *et al*. 1993). The length of microsatellite fragments reflects insertion/deletion mutations. Methods such as inter-simple sequence repeats (ISSR), which measure the distance between microsatellites were based on SSR. Similarly to these based on microsatellites, there are markers that determine the retrotransposon positions (Flavell *et al*. 1999) and the distances between them (Kalendar *et al*. 1999). Microsatellite- and transposon-based methods are still in use as they are cheap, however, labour intensive and require prior development. Lastly, there are sequence-based markers genotyped through Sanger sequencing (Sanger & Coulson 1975). In these cases polymorphism is investigated directly via analysis of nucleotide codes and mutations. This method is expensive,

however, it identifies many single nucleotide polymorphisms (SNPs), and each of these constitutes a separate marker.

1.5.2 High-throughput molecular markers

The robustness of population genetics analyses increases with number of studied loci and individuals. There are multiple methods that allow genotyping large number of loci using conventional tools; for example short sequence repeats (SSR) analysis. However, these are very laborious when a very high number of loci are required. Alternatively, next-generation sequencing (NGS) platforms with use of genotyping by sequencing (GBS) techniques offer marker discovery and genotyping of large numbers of loci at a low cost. Sequencing whole genomes is not necessary for phylogeographic and population genetics studies. A subset of a genome is sufficient to answer questions about relationship between individuals or populations. At the same time reduction of genotyping data will dramatically increase computational speed of otherwise intensive statistical analyses. Furthermore, by sequencing only a subset of a genome one resolves problems with genome assembly for non-model organisms.

Genome sampling can be carried out by multiplex PCR reactions, targeted enrichment with probes or by restriction reactions. The multiplex PCR method assumes amplification of genomic regions of interest, which were earlier found in Sanger sequencing experiments. It is essentially a parallelization of conventional sequencing approach and therefore is often called parallel tagged sequencing (Meyer *et al*. 2008). Alternatively, one can enrich genomic DNA with regions of interest before library preparation with hybridization probes (Mamanova *et al*. 2009). This approach is referred to as sequence capture. Hybridization can be carried out in solution or on an array. Prior knowledge about sequences of interesting genomic regions is necessary to execute multiplex PCR or DNA capture approach. For this reason, genome-scale population genetics approaches were restricted to model organisms only. Flax genome is still being reconstructed and therefore it would be very difficult to apply PCR and/or DNA capture approaches for marker development. Although its genome has been sequenced (Wang et al. 2012), the quality of assembly and its annotation is poor. However, there are new genome sampling techniques, which allow sequencing multiple homologous genomic regions for individuals of

non-model species without a reference genome or annotation of partial genome sequences.

Restriction-site associated DNA sequencing (RADseq) (Baird *et al.* 2008) with *de novo* assembly option give the opportunity to sequence multiple loci for hundreds of individuals even if they are in non-model organisms (Davey & Blaxter 2011). In this method, flanking regions of restriction sites are sequenced. By choosing appropriate restriction enzyme one can manipulate number of genotyped loci in genome. RADseq was used multiple times in plants, with the aim of: 1) genotyping for QTL mapping in barley (*Hordeum vulgare*) and perennial ryegrass (*Lolium perenne)* (Chutimanitsakun et al. 2011; Pfender et al. 2011), 2) SNP discovery in eggplant (*Solanum melongena*) (Barchi *et al.* 2011; Barchi *et al.* 2012), cardoon (*Cynara cardunculus*) (Scaglione *et al.* 2012), birch (*Betula nana*) (Wang *et al.* 2013) and rapeseed (*Brassica napus*) (Bus *et al.* 2012), 3) marker-associated molecular breeding in blue lupin (*Lupinus angustifolius*) (Yang *et al.* 2012) and 4) construction of high-density genetic map in grape (*Vitis vinifera*) (Wang *et al.* 2012b). Interestingly, to the best of this author's knowledge, RADseq has not yet been used in plant population genetics studies. It was used to measure gene flow between poplars on the species- rather than population-level (Stoelting *et al.* 2013). Authors noted that this method allows investigating the divergence between species that are interfertile. Hence, RADseq could be successfully used to study divergence between cultivated species and their wild ancestors. The above evidence suggest that RADseq is a good method to deliver data that will help in tackling questions posed in this thesis about post-domestication gene flow between the two non-model species: pale and cultivated flax.

1.5.3 Non-neutral markers in investigation of domestication and crop evolution

Non-neutral markers are associated with loci under selection. These loci usually are under positive selection when they exhibit adaptive properties. Adaptive evolution is driven by the selection forces, which favour certain alleles in the population based on their impact on fitness in current environment. When an allele is favoured, the phenomenon is referred to as positive selection, whereas, when an allele is eliminated negative selection is inferred. Existence of either types of selection leads to a dynamic change in population's allele frequency. The rates of this change

depend on selection type and strength. If a single allele is conferring the highest fitness we observe the directional selection. By contrast, when heterozygotes have the highest fitness we might expect the balancing selection. Lastly, when two or more alleles are favoured simultaneously we observe disruptive selection. Selection strength is influenced by the fitness of allele carriers in current environment. That means that the higher the difference in fitness between the phenotypes the stronger selection applies. If a population sustains the selection it adapts successfully, if the opposite is true, it perishes.

Genes under selection in domesticated species are of particular interest to study different stages of crop evolution and adaptation. There are three recognized stages of crop evolution, each characterized by different set of changes in crops' morphology and phenology (Brown *et al.* 2009). Firstly, domestication, which is associated with introducing domestication syndrome traits into crops. Secondly, spread of crop plants and their adaptation to local environments that is usually linked to a change in phenology traits such as germinating and flowering time. Finally, increased yield reflects the stage of crop improvement and specialization. All changes in crop morphology and phenology are recognized as being under positive selection. However, they have distinct effect on molecular variation within a population. Domestication syndrome traits were under selection at the whole species level and usually resulted in low levels of nucleotide variation in general (Smith & Haigh 2007) and/or an unusual number of rare substitutions (Tajima 1989). This is a result of purifying selection, where new alleles appear but they never reach high frequencies due to high preference for the main allele. Purifying selection could also occur on local scale, leading to local adaptations. In this case however, it creates local strong alleles. The signature it leaves in genetic data is characterized by reduced effective migration on the subpopulation level and in consequence - strong population structure (Charlesworth *et al.* 1997). The difference in selection signatures in molecular diversity data can be used to distinguish domestication syndrome from local adaptation traits.

1.5.4 A locus-specific approach to the adaptive evolution of flax

To find signature of selection linked to the emergence of fibre flax molecular markers associated with this particular purpose of cultivation are needed. To the best

of the author's knowledge there are no genetic markers in flax that would enable unambiguous discrimination of oil and fibre varieties. Genes that are associated with oil production such as *SAD2* locus (Allaby *et al.* 2005; Fu *et al.* 2012) did not allow for accurate identification of oil varieties. Molecular markers specific to fibre varieties are scarce and unreliable. At the time of writing, the list of genes that might be under disruptive selection between fibre and oil flax has been published (Soto-Cerda *et al.* 2013). These genes might prove very effective in tracing the origins of fibre flax.

A different approach might be undertaken to trace emergence of fibre varieties. Based on archaeological data (see Section ***1.3.1***) it is assumed that the specialized fibre varieties arose in the Central Europe, probably as summer varieties. Therefore there might be a genetic link between the population's signature of adaptation to high latitude, summer varieties and fibre cultivars. Furthermore, based on archaeological dating emergence of fibre flax varieties was likely the result of local adaptation rather than domestication. Therefore, one would expect to see strong population structure and more than one strong allele in locus associated with latitudinal adaptation. Latitudinal adaptation is associated with changes in flowering time (see Section ***1.3.2***). Hence, non-neutral markers should be developed in genes responsible for flowering to investigate latitudinal adaptation and emergence of flax fibre varieties. The most robust markers are whole sequences of genes, which are genotyped by re-sequencing. The disadvantage of this approach is its cost. However, this method allows not only differentiating genetic subpopulations, but also additionally investigating the molecular mechanism of adaptation and hence is best suited to address the questions of this project.

1.5.5 Signature of selection in genome-wide SNP data

By contrast to the deductive (top-down) approach presented above, it is possible to use inductive (bottom-up) approach to studying adaptive evolution. Recent studies using genome-wide SNPs show that GBS methods are excellent in identifying the genomic regions with signature of selection. Regions with increased levels of linkage disequilibrium (LDE), unusual levels of nucleotide variability, increased levels of population differentiation, or skewed allele frequency spectra may indicate the presence of loci under selection. Recently the use of increased population

subdivision signal, as measured by fixation index (F_{ST}), is a common practice in searching for domestication syndrome candidate genes and genomic regions that might be linked to local adaptations of crop plants (Walsh 2008). For instance, in maize, a genomic scan for selection signatures was used to identify genes that were involved in the domestication and improvement stages of evolution (Yamasaki *et al.* 2005). Furthermore, in sunflower, multiple loci under selection were found when landraces were compared to modern cultivars, some of them where located in regions inhabited by important QTL's (Chapman *et al.* 2008). This approach proves extremely effective in investigating post-domestication evolution of crop plants. It would be most robust to carry out such analyses based on sequences of whole genomes, however, this approach is very expensive and computationally demanding. Instead, the subset of genomic data could be used. Sequencing methods such as RADseq prove to be very well suited for this purpose.

1.5.6 Neutral molecular markers in population genetics

Neutral molecular markers can be used to characterise the genetic diversity of species and are thus extremely useful tools in studying population structure, demography, migrations and evolution of crop plants. Mutations are stochastic and thus there is a chance for a DNA substitution in any individual of particular subpopulation. The mutation could increase in frequency in this subpopulation through genetic drift and become a differentiating agent from other subpopulations; this mutation is referred to as private mutation. In this way, polymorphism and frequency of linked alleles can be used to establish population structure. However, because individuals within one species (even if they are from different populations) are interfertile, we observe migration events that disrupt strict population structure. In natural environment the migration is dependent on geographic distances between these populations. In simple example, new derived allele from population A could be transferred to population B via hybridization. In this case we would expect the frequencies of this allele in population B to be dependent on number of migration events. There are more of these events with shorter distance and less isolation barriers between populations. One simple method of quantifying population separation is based on so called F-statistics (Wright 1950). The main idea behind it is to measure reduction in heterozygosity when compared to expected value given by the Hardy-Weinberg equilibrium.

There were numerous attempts to study population structure, phylogenetics and phylogeography of flax species based on neutral markers. The focus of most of these studies was cultivated flax. Its diversity and population structure was investigated based on RAPD, AFLP, IRAP and SSR markers. RAPD markers did not allow dicriminating main four flax convarieties (Diederichsen & Fu 2006; Everaert *et al.* 2001). Based on AFLP markers, discrimination between fibre and oil flax was possible, however, only PCA analysis in contrast to phylogenetic approach enabled discrimination (Vromans 2006). In IRAP study, it has been observed that genetic diversity of oil and fibre varieties is largely overlapping (Smykal *et al.* 2011). Phylogeographic analyses based on SSR markers revealed population division between South American, North American and Eastern European flax (Soto-Cerda & Cloutier 2013; Soto-Cerda *et al*. 2013; Soto-Cerda *et al*. 2012). There are only a few projects that included pale flax in analyses of flax diveristy and investigation of flax domestication. Studies based on ISSR, EST and resequencing resulted in resolving relationship between pale flax, indehiscent and winter varieties (Fu & Peterson 2010; Uysal *et al.* 2010). Authors of chloroplast loci resequencing project conclude that the relationship between pale and cultivated flax could be more complex and post-domestication gene flow possible (Fu & Allaby 2010).

For purpose of in-depth analyses of relationship between pale and cultivated flax, a large number of markers have to be developed. GBS with use of new generation sequencing is an ideal platform to deliver molecular markers. Since flax is not a model plant and its genome has not been yet reconstructed use of RADseq approach is favoured. It will allow developing SNP markers in both species of flax and simulatonously genotype them for large number of samples. This approach has not yet been used to develop and genotype markers in flax, however, was very successful in other plant species (reviewed in Section ***1.5.2***).

1.5.7 Review of population genetics analyses

The choice of statistical analyses for molecular diversity data depends on the number of genotyped samples, type of molecular data and biological question. For investigation of relationship between individuals we can execute clustering analyses. They require relatively few samples, but the power of such analyses is proportional to the number of molecular markers. It allows samples to be hierarchically grouped

into the form of a clustering tree, which may serve as a proxy of the relationship between individuals or subpopulations (tips of the tree). This approach is very useful for handling large datasets of simple markers such as RAPD, AFLP or SNPs. Each pair of nodes can be linked with only one branch and therefore this analysis does not show genetic recombination, hybridization, or linage sorting effects. These phenomena could be investigated with use of clustering networks such as *SPLITSTREE*. In this analysis, nodes can be linked with multiple branches and with multiple other nodes, allowing more complex relationships within a population to be observed. Full sequence data are better suited for clustering network analyses.

A phylogenetic analysis provides a more sophisticated statistical approach to investigate the relationship of genes or species through phylogenetic analyses. In this instance, full sequence data are preferred because the estimation of evolution models and substitution rates is based on both variable and invariable characters. Phylogenetic trees based on one locus, such as a gene, represent the evolution and divergences of this locus. In some cases, when each species contains a different haplotype in this locus we can infer the species tree based on the gene tree. Loci that evolve neutrally (especially chloroplast loci) are usually good candidates for reconstructing phylogeny of species. The most common methods of phylogenetic tree estimation are Maximum Parsimony (MP), Maximum Likelihood (ML) and Bayesian Inference (BI). The latter is widely used because of the robustness of Bayesian method and because it is considered less computationally demanding than ML. A variant of this method uses a coalescent genealogy sampler (*BEAST*) to estimate the time of gene or population divergence using molecular clock approach.

The analyses described above can estimate the relationship between populations or individuals, but they cannot assign an individual to population and infer subpopulation structure. There are two types of methods used for the purpose of structuring population data. These methods could be divided into methods based on admixture models and methods based on principle component analysis (PCA) (Engelhardt & Stephens 2010). In the former, the number of K populations is estimated. Each individual is assumed to have inherited a proportion of its ancestry (admixture) from each of the K populations. This approach is used in *STRUCTURE*, *INSTRUCT* and *BAPS* software. In PCA, individuals are mapped onto a two-dimensional space in a way that directly reflects the genetic similarity between them.

Both methods require a large number of samples and are more robust when a large dataset of molecular markers are used. While the admixture model approach better reflects the genetic migration between populations, the PCA can demonstrate correlation between genetic and geographic distance particularly well.

Analyses of demographic and migration models based on genetic data help us to understand the past dynamics and interactions between subpopulations. These approaches test the likelihoods of different scenarios based on allele frequency spectra (*δaδi*) or coalescence (*IMa2*, *MIGRATE-N*). In these scenarios, parameters such as the rate of gene flow, the rate of population expansion and the time of subpopulation divergence are taken into account. These approaches require a high number of individuals and molecular markers. Ideally, these markers should not be in linkage-disequilibrium, otherwise results of analyses could be biased. *MIGRATE-N* allows the user to test different migration models and to estimate the number of genetic migrants per generation between many subpopulations. *IMa2* is limited to only two subpopulations but has an advantage in that it allows estimating the time of divergence and significant gene flow.

1.6 OUTLINE OF THE THESIS

In **Chapter 2**, the assembly of plant material that is necessary to conduct phylogeographic and population analyses is described. In **Chapter 3**, genes responsible for control over flowering time in flax have been investigated. Paralogs belonging to *FT*/*TFL* gene family have been identified in flax. Structure and active centre of these genes was compared to orthologs in model organisms and their phylogeny was reconstructed. In **Chapter 4**, two loci from *FT*/*TFL* family, *LuTFL1* and *LuTFL2* were chosen for re-sequencing and population genetics analyses. In short, *LuTFL1* haplotype network provides evidence for post-domestication gene flow from pale to cultivated flax while population parameters suggest signature of selection within cultivated flax gene pool. Additionally, *LuTFL1* haplotypes are correlated with latitude. In **Chapter 5**, flowering time, growth determinacy, stem height and floral architecture of flax and its relation to latitude was modelled based on studies in other organisms. This model was tested using data collected for cultivated flax in PGRC and data for pale flax. The latter included morphological, phenological observations and expression analyses. In **Chapter 6**, next generation sequencing data from RADseq approach are utilized to infer about population structure in both cultivated and pale flax. Results revealed that there is latitudinal gradient in fibre flax varieties, while oil varieties are not structured. The RADseq data is used to quantify levels of genetic migration between geographically different populations of pale and cultivated flax. The results confirm that domestication occurred in the Near East region, however, high levels of post-domestication gene flow is recorded in Europe. Finally in **Chapter 7**, conclusions from this study are discussed in light of findings in other crop species. The null hypothesis finds its support in this project, however, transgenic experiments should be carried out in future to improve the evidence.

CHAPTER 2: ASSEMBLY OF PLANT MATERIAL

2.1 INTRODUCTION

A large number of accessions covering the Near East, Asia Minor and Europe is required to study flax introduction to the Northerly latitudes. Therefore assembly of appropriate plant material is crucial for this project. Flax specimens are stored in *ex situ* collections around the world. Living material is stored in gene banks in the form of seeds, while desiccated material can be found in herbaria and museums. Gene bank collections are very well described. There are 48,282 accessions of cultivated flax available in the 33 largest collections, however, only 10,000 of these are unique (Diederichsen 2007). An effort was made to assemble a core collection of flax at Plant Genetic Resources of Canada (Saskatoon). It resulted with a panel of 381 accessions, representing flax diversity (Diederichsen *et al*. 2012). The wild relative of flax (pale flax) is underrepresented in the seed collections. There are only 279 accessions available worldwide and these poorly cover the Near East, Asia Minor and the Balkans (Diederichsen 2007). This issue was partially resolved during a 2007 collection trip to Turkey, in which 34 populations of pale flax were sampled (Uysal et al. 2012). Further sampling is required to cover the Balkan countries. The herbaria resources are vast and hence usually difficult to catalogue. There are no estimates of the number of flax specimens stored in herbaria. The largest herbarium collections are listed in *index herbariorum* (Holmgren & Holmgren 1998). Additional, non-herbaria, historic resources are available in museum collections such as the Economic Botany Collection at the Royal Botanic Gardens, Kew.

The aim of this chapter is to describe the assembly of cultivated and pale flax accessions, which were used in this project. Plant material was collected from seed banks, museums, herbaria and sampled from pale flax wild habitats in the Balkans.

2.2 MATERIALS AND METHODS

A request for seeds of modern cultivars, historic landraces and wild populations of pale flax was sent to PGRC (Saskatoon) and the Vavilov Institute (St. Petersburg). Representative accessions for European intermediate, oil and fibre varieties were requested. Sampling of herbarium material was carried out at hosting institutions using sterile tools and gloves to avoid cross-contamination. Where possible, capsules were sampled; otherwise, pieces of stem and leaves were detached and moved to sterile 1.5 ml tubes. Historic seeds from botanic museums were sampled using sterile tools and isolated in 1.5 ml tubes. Sampling of historic material was conducted under destructive sampling agreements with hosting institutions.

A collection trip was organized in order to sample pale flax populations from south-eastern Europe. The following countries were visited: Croatia, Montenegro, Albania, Greece and Bulgaria. Locations of pale flax habitats were retrieved from local databases or communicated via e-mail by local scientists (Angelova 2011; Nikolic 2010; Strid 2011). Further locations were found during exploration of coastal and mountain regions. Sampling locations were described with following information: GPS coordinates, altitude, approximate location relative to characteristic features of landscape, soil type and botanical habitat. Leaf tissue and seed capsules were collected from at least three different pale flax gatherings, which were separated by at least 40 m. Material was sealed in separate envelopes with small amount of silica gel. A typical plant was selected and its morphology described, including information such as: plant height, capsule size, number of stems, and proportion of stem that include floral branches. Furthermore, photographs of a typical plant were taken.

2.3 RESULTS

Samples of cultivated flax used in this project consisted of historic material, landraces and modern cultivars. Most of the plants in these categories were representative of fibre, oil, intermediate and dehiscent varieties, however, they did not always match morphological criteria for respective convarieties suggested by Kulpa & Danert (1967). Seed material was sampled from gene bank accessions (Table **2.1**). All of the modern varieties (58 samples) and 18 samples of landraces were contributed by collaborators at Plant Genetic Resources of Canada, Agriculture and Agri-Food Canada, Saskatoon Research Centre. This material is covered by The International Treaty on Plant Genetic Resources for Food and Agriculture. Seeds were shipped under a Standard Material Transfer agreement. Unique samples of 28 historic landrace accessions were obtained from Plant Genetic Resources of the N. I. Vavilov Research Institute of Plant Industry, St. Petersburg, Russia. Transfer of seed material was organized by collaborators from NordGen, Nordic Genetic Resources, Alnarp, Sweden.

Table 2.1: The list of seed samples of cultivated flax

Name	Source	ID No.	Country of origin	Convariety	Status
M001	PGR Canada	98833	Russia	Crepitans	Dehiscent
M002	PGR Canada	100837	Turkey	Crepitans	Dehiscent
M003	PGR Canada	100852	Portugal	Crepitans	Dehiscent
M004	PGR Canada	97606	Spain	Crepitans	Dehiscent
M005	PGR Canada	98507	Israel	Usitatissimum	Fibre var.
M006	PGR Canada	98511	Israel	Usitatissimum	Fibre var.
M007	PGR Canada	97048	Iran	Usitatissimum	Fibre var.
M008	PGR Canada	97180	Iran	Elongatum	Fibre var.
M009	PGR Canada	98162	Iran	Usitatissimum	Fibre var.
M010	PGR Canada	98561	Turkey	Usitatissimum	Fibre var.
M011	PGR Canada	98562	Turkey	Usitatissimum	Fibre var.
M012	PGR Canada	96889	Turkey	Usitatissimum	Fibre var.
M013	PGR Canada	98299	Hungary	Usitatissimum	Fibre var.
M014	PGR Canada	98300	Hungary	Usitatissimum	Fibre var.
M015	PGR Canada	98303	Hungary	Usitatissimum	Fibre var.
M016	PGR Canada	98475	Germany	Elongatum	Fibre var.
M017	PGR Canada	18978	Germany	Elongatum	Fibre var.
M018	PGR Canada	18991	Poland	Elongatum	Fibre var.
M019	PGR Canada	97325	Poland	Usitatissimum	Fibre var.
M020	PGR Canada	97326	Poland	Usitatissimum	Fibre var.
M021	PGR Canada	18990	Czech Republic	Usitatissimum	Fibre var.
M022	PGR Canada	98700	Czech Republic	Elongatum	Fibre var.
M023	PGR Canada	98479	Czech Republic	Usitatissimum	Fibre var.
M024	PGR Canada	18995	Russia	Elongatum	Fibre var.
M025	PGR Canada	32544	Russia	Elongatum	Fibre var.
M026	PGR Canada	101099	Russia	Elongatum	Fibre var.
M027	PGR Canada	101056	Ukraine	Usitatissimum	Fibre var.
M028	PGR Canada	30862	Ukraine	Usitatissimum	Fibre var.
M029	PGR Canada	101020	Kazakhstan	Usitatissimum	Fibre var.
M030	PGR Canada	18988	France	Elongatum	Fibre var.
M031	PGR Canada	18982	France	Usitatissimum	Fibre var.
M032	PGR Canada	18989	France	Usitatissimum	Fibre var.
M033	PGR Canada	97043	Spain	Usitatissimum	Fibre var.
M034	PGR Canada	101031	Morocco	Usitatissimum	Fibre var.
M035	PGR Canada	101025	Morocco	Usitatissimum	Fibre var.
M036	PGR Canada	19011	Tunisia	Usitatissimum	Fibre var.
M037	PGR Canada	100960	Tunisia	Usitatissimum	Fibre var.

M038	PGR Canada	18999	The Netherlands	Usitatissimum	Fibre var.
M039	PGR Canada	100930	The Netherlands	Elongatum	Fibre var.
M040	PGR Canada	100929	The Netherlands	Elongatum	Fibre var.
M041	PGR Canada	100909	Palestine	Usitatissimum	Landrace
M042	PGR Canada	100893	Iran	Usitatissimum	Landrace
M043	PGR Canada	113611	Turkey	-	Landrace
M044	PGR Canada	113613	Turkey	-	Landrace
M045	PGR Canada	101035	Russia	Elongatum	Landrace
M046	PGR Canada	101046	Russia,	Usitatissimum	Landrace
M047	PGR Canada	100922	Malta	Usitatissimum	Landrace
M048	PGR Canada	100896	Egypt	Usitatissimum	Landrace
M049	PGR Canada	101028	Tunisia	Usitatissimum	Landrace
M050	PGR Canada	98163	Iran	Usitatissimum	Oil var.
M051	PGR Canada	98164	Iran	Usitatissimum	Oil var.
M052	PGR Canada	113643	Turkey	-	Oil var.
M053	PGR Canada	97017	Turkey	Usitatissimum	Oil var.
M054	PGR Canada	98987	Italy	-	Oil var.
M055	PGR Canada	98988	Italy	-	Oil var.
M056	PGR Canada	97321	Romania	Usitatissimum	Oil var.
M057	PGR Canada	97322	Romania	Usitatissimum	Oil var.
M058	PGR Canada	18992	Hungary	Usitatissimum	Oil var.
M059	PGR Canada	98278	Hungary	Usitatissimum	Oil var.
M060	PGR Canada	97758	Germany	Usitatissimum	Oil var.
M061	PGR Canada	97760	Germany	Usitatissimum	Oil var.
M062	PGR Canada	97332	Poland	Usitatissimum	Oil var.
M063	PGR Canada	97176	Czech Republic	Usitatissimum	Oil var.
M064	PGR Canada	97344	Czech Republic	Usitatissimum	Oil var.
M065	PGR Canada	18996	France	Mediterraneum	Oil var.
M066	PGR Canada	98752	France	Usitatissimum	Oil var.
M067	PGR Canada	98522	UK	Usitatissimum	Oil var.
M068	PGR Canada	97044	Spain	Usitatissimum	Oil var.
M069	PGR Canada	98826	Egypt	Usitatissimum	Oil var.
M070	PGR Canada	97439	Egypt	Usitatissimum	Oil var.
M071	PGR Canada	98880	Morocco	Usitatissimum	Oil var.
M072	PGR Canada	30852	Russia, Rostov	Usitatissimum	Wild
M073	PGR Canada	30841	Kazakhstan	Usitatissimum	Wild
M074	PGR Canada	30844	Ukraine	Usitatissimum	Wild
M075	PGR Canada	30846	Ukraine	Usitatissimum	Wild
M076	PGR Canada	30842	Kazakhstan	Usitatissimum	Wild
H080	VIR Russia	562	Russia		
H081	VIR Russia	725	Russia		
H082	VIR Russia	730	Russia		
H083	VIR Russia	776	Belarus	Fibre	Landrace
H084	VIR Russia	791	Belarus	Fibre	Landrace
H085	VIR Russia	889	Ukraine	Intermediate	Landrace
H086	VIR Russia	900	Italy	Intermediate	Landrace
H087	VIR Russia	901	Ukraine	Intermediate	Landrace
H088	VIR Russia	925	Armenia	Oil	Landrace
H089	VIR Russia	929	Kazakhstan	Oil	Landrace
H090	VIR Russia	1028	Russia	Fibre	Landrace
H091	VIR Russia	1032	Russia	Fibre	Landrace
H092	VIR Russia	1042	Belarus	Fibre	Landrace
H093	VIR Russia	1693	Turkey	Oil	Landrace
H094	VIR Russia	1723	Turkey	Oil	Landrace
H095	VIR Russia	1727	Turkey	Oil	Landrace
H096	VIR Russia	1859	Iran	Oil	Landrace
H097	VIR Russia	1930	Argentina	Intermediate	Landrace
H098	VIR Russia	1931	Tunisia	Intermediate	Landrace
H099	VIR Russia	1933	Argentina	Intermediate	Landrace
H100	VIR Russia	2458	Turkey	Oil	Landrace
H101	VIR Russia	2526	Italy	Intermediate	Landrace
H102	VIR Russia	3688	Latvia	Fibre	Landrace
H103	VIR Russia	17	Russia	Fibre	Historic var.
H104	VIR Russia	193	The Netherlands	Fibre	Historic var.
H105	VIR Russia	253	The Netherlands	Fibre	Historic var.
H106	VIR Russia	462	Turkey	Oil	Historic var.
H107	VIR Russia	1114	Israel	Intermediate	Historic var.
H108	VIR Russia	3683	Latvia	Fibre	Historic var.
H109	VIR Russia	3933	Germany	Fibre	Historic var.
H110	VIR Russia	4165	France	Intermediate	Historic var.
H111	VIR Russia	5339	Ukraine	Intermediate	Historic var.
H112	VIR Russia	5546	Yugoslavia	Intermediate	Historic var.
H113	VIR Russia	5555	Yugoslavia	Intermediate	Historic var.

Table 2.2: The list of herbarium and historic samples of cultivated flax

Name	Source	ID No.	Country of origin	Convariety	Status
H001	UoW Poland	HPW02	Poland	?	1909
H002	UoW Poland	HPW03	Poland	?	1968
H003	UoW Poland	HPW04	Poland	?	1826-1895
H004	UoW Poland	HPW05	Poland	?	1826-1896
H005	UoW Poland	HPW06	Poland	?	1869
H006	UoW Poland	HPW08	Lithuania	?	1899
H007	UoW Poland	HPW11	Belarus	?	?
H008	UoW Poland	HPW12	Poland	?	1903
H009	UoW Poland	HPW13	Poland	?	1821
H010	UoW Poland	HPW14	Belarus	?	1902
H011	UoW Poland	HPW15	Poland	?	1891
H012	UoW Poland	HPW16	Poland	?	1963
H013	UoW Poland	HPW17	Poland	?	1961
H014	UoW Poland	HPW18	Germany	?	1896
H015	UoW Poland	HPW19	Germany	?	1986
H016	UoW Poland	HPW20	Ukraine	?	1915
H017	UoW Poland	HPW21	Austria	?	1869
H018	UoW Poland	HPW22	Sweden	?	1920
H019	UoW Poland	HPW23	Italy	?	1896
H020	UoW Poland	HPW24	Ukraine	?	1912
H021	UoW Poland	HPW25	Russia	?	1916
H022	UoW Poland	HPW26	Romania	?	1903
H023	UoW Poland	HPW27	Serbia	?	1899
H024	UoW Poland	HPW28	Portugal	?	1888
H025	UoO UK	HUO01	UK	?	1895
H026	UoO UK	HUO02	UK	?	1897
H027	UoO UK	HUO03	UK	?	1892
H028	UoO UK	HUO04	UK	?	1909
H029	UoO UK	HUO05	UK	?	1892
H030	UoO UK	HUO06	UK	?	1892
H031	UoO UK	HUO07	UK	?	1875
H032	UoO UK	HUO08	UK	?	1873
H033	UoO UK	HUO09	Dalmaka	?	1914
H034	UoO UK	HUO11	UK	?	1640-1683
H035	UoO UK	HUO13	Greece	?	1780-1790
H036	NHM UK	HUNG01	Sweden	?	1890
H037	NHM UK	HUNG02	Sweden	?	1870
H038	NHM UK	HUNG03	Macedonia	?	<1877
H039	NHM UK	HUNG04	Greece	?	1934
H040	NHM UK	HUNG05	Germany	?	1829
H041	NHM UK	HUNG06	Italy	?	1888
H042	NHM UK	HUNG07	Macedonia	?	1918
H043	NHM UK	HUNG08	Denmark	?	1942
H044	NHM UK	HUNG09	Switzerland	?	1861
H045	NHM UK	HUNG10	France	?	1813
H046	NHM UK	HUNG11	France	?	1923
H047	NHM UK	HUNG12	Germany	?	1950
H048	NHM UK	HUNG13	France	?	1872
H049	NHM UK	HUNG32	Iran	?	1845
H050	NHM UK	HUNG33	Turkey	?	1957
H051	NHM UK	HUNG34	Syria	?	1945
H052	NHM UK	HUNG01	UK	?	1871
H053	NHM UK	HUNB02	UK	?	?
H054	NHM UK	HUNB03	UK	?	1837
H055	NHM UK	HUNB04	UK	?	1881
H056	NHM UK	HUNB05	UK	?	1846
H057	NHM UK	HUNB06	UK	Crepitans	1909
H058	NHM UK	HUNB07	UK	?	1891
H059	NHM UK	HUNB08	UK	?	1904
H060	RBGK UK	HUK01	Ireland	Oil	?
H061	RBGK UK	HUK02	The Netherlands	Oil	<1929
H062	RBGK UK	HUK03	Italy	Oil	<1852
H063	RBGK UK	HUK04	Russia	Oil	<1878
H064	RBGK UK	HUK05	Russia	Oil	<1878
H065	RBGK UK	HUK06	Turkey	Oil	1856
H066	RBGK UK	HUK07	Morocco	Oil	<1928
H067	RBGK UK	HUK08	Tunisia	Oil	<1908
H068	RBGK UK	HUK09	Russia	Oil	<1878
H069	RBGK UK	HUK10	Russia	Oil	<1852
H070	RBGK UK	HUK11	Egypt	Oil	<1853
H071	RBGK UK	HUK13	Baltic	Oil	1899
H072	RBGK UK	HUK14	Russia	Oil	1899
H073	RBGK UK	HUK15	Ukraine	Oil	1899
H074	RBGK UK	HUK16	UK	Oil	?
H075	RBGK UK	HUK17	The Netherlands	Oil	?
H076	RBGK UK	HUK18	Greece	Oil	?
H077	RBGK UK	HUK20	UK	Oil	1843
H078	RBGK UK	HUK21	Iran	Oil	?
H079	RBGK UK	HUK22	Palestine	Oil	?

Table 2.3: Description of samples of wild flax and their geographic coordinates

ID	Species	Country	Coordinates	Altitude	Soil type	Habitat
W63	*L. tenuifolium*	Croatia	N44°33'531 E015°45'095	685 m	Rocky, gravel-silt, fresh, on limestone	Dry grassland: grass, *Achillea*, *Potentilla*, *Echium*, *Thymus*
W64	*L. tenuifolium*	Croatia	N43°40'172 E015°56'275	38 m	Rocky, silt, dry, brown, on limestone	Dry grassland: grass, *Pinus*, *Allium*
W65	*L. bienne*	Croatia	N43°54'555 E016°27'880	358 m	Rocky, clay, fresh, grey-brown, on limestone	Dry grassland: grass, *Populus*, *Potentilla*, *Rubus*
W66	*L. bienne*	Croatia	N43°28'854 E016°50'565	250 m	Clay, fresh, brown, on limestone	Dry grassland: grass, *Salix*, *Potentilla*, *Sium*, *Convonvulus*, *Anethum*
W67	*L. bienne*	Croatia	N43°01'265 E017°27'060	1 m	Sandy, fresh,	Reeds: grass, *Phragmites*, *Convolvulus*, *Juncus*, *Daucus*
W68	*L. bienne*	Montenegro	N42°14'799 E018°53'767	10 m	Rocky, on limestone	Dry grassland: grass, *Juniperus*, *Linum*, *Laurus*,
W69	*L. bienne*	Albania	N41°10'887	2 m	Sandy, dry, ruderal	Dry grassland: grass, *Apium Daucus*, *Cichorium*, *Nerium*
W70	*L. bienne*	Greece	N39°54'514 E020°22'296	281 m	Rocky, on limestone	Forest clearings: grass, *Platanus*, *Tilia*, *Quercus*, *Rubus*, ferns
W71	*L. strictum*	Greece	N39°33'939 E021°00'222	991 m	Rocky, mountain soil, grey, on limestone	Dry grassland: grass, *Juniperus Ilex*, *Phragmites*
W72	*L. bienne*	Greece	N38°59'438 E021°09'233	1 m	Rocky, gravel-silt, fresh, on limestone	Dry grassland: grass, *Achillea*, *Potentilla*, *Echium*, *Thymus*
W73	*L. flavum*	Greece	N38°56'082 E021°51'910	1480 m	Rocky, mountain soil, grey, on limestone	Dry grassland: grass, *Heracleum*
W74	*L. bienne*	Greece	N39°03'265 E021°52'565	946 m	Rocky, mountain soil, grey, on limestone	Dry grassland: grass, Achillea, *Potentilla*, *Echium*, *Thymus*
W75	*?*	Greece	N39°43'351 E021°38'774	548 m	Rocky, on limestone	Dry grassland: grass, *Platanus*
W76	*L. bienne*	Greece	N39°47'476 E021°38'599	621 m	Rocky, clay, dry, on limestone	Dry grassland: grass, *Linum*, *Quercus*
W77	*L. bienne*	Greece	N39°58'435 E021°30'800	835 m	Rocky, clay, dry, brown, on limestone	Dry grassland: grass, *Quercus*, *Centaurea*, *Daucus*
W78	*?*	Greece	N40°14'761 E021°54'100	390 m	Rural, on limestone	Dry grassland: grass, empty fields
W79	*L. austriacum*	Greece	N40°44'946 E021°41'511	581 m	Rocky, sandy-gravel, fresh, brown, on limestone	Dry grassland: grass
W80	*?*	Greece	N40°32'090 E023°06'895	288 m	Rocky, on limestone	Dry grassland: grass, *Pinus*, *Quercus*, *Juniperus*
W81	*L. bienne*	Greece	N40°21'625 E023°56'005	2 m	Sandy, marine	Dry grassland: grass, Foeniculum, *Convonvulus*, *Cirsium*
W82	*L. bienne*	Greece	N40°35'028 E023°47'074	129 m	Rocky, gravel, dry, on limestone	Dry grassland: grass, *Rododendron*, *Rubus*, *Platanum*
W83	*?*	Greece	N40°52'476 E023°34'349	210 m	Rocky, black, dry, on limestone	Dry grassland: grass, *Quercus*, *Helianthus*
W84	*L. tenuifolium*	Bulgaria	N41°54'350 E023°18'375	924 m	Rocky, silt, dry, on limestone	Dry grassland: grass, *Pinus*, *Prunus*, *Cirsium*
W85	*L. bienne*	Bulgaria	N41°35'840 E023°43'843	591 m	Rocky, silt, dry, yellow, on limestone	Dry grassland: grass, *Pinus*
W86	*L. decumbens*	Bulgaria	N41°33'370 E024°01'651	865 m	Rocky, sandy-silt, fresh, on limestone	Dry grassland: grass, *Salix*, *Prunus*
W87	*L. austriacum*	Bulgaria	N42°04'545 E026°18'851	357 m	Rocky, gravel-clay, fresh, brown, on limestone	Dry grassland: grass, *Rosa*, *Rubus*, *Prunus*, *Daucus*, *Berteroa*
W88	*?*	Bulgaria	N42°43'100 E027°45'315	89 m	Rocky, gravel-silt, fresh, on limestone	Dry grassland: grass,
W89	*L. tenuifolium*	Bulgaria	N42°44'695 E027°44'392	345 m	Rocky, on limestone	Dry grassland: grass,
W90	*L. austriacum*	Bulgaria	N43°21'38 E028°27'57	40 m	no info	no info
W91	*L. austriacum*	Bulgaria	N43°17'173 E027°18'452	217 m	Rocky, clay, fresh, grey, on limestone	Dry grassland: grass, *Prunus*, *Vicia*, *Daucus*, *Linum*, *Crataegus*
W92	*L. tenuifolium*	Bulgaria	N43°25'835 E025°08'452	96 m	Rocky, on limestone	Dry grassland: grass,
W93	*L. catharticum*	Romania	N45°01'157 E021°54'288	520 m	Rocky, fresh, on limestone	Dry grassland: grass,
W94	*L. bienne*	Croatia	N45°08'308 E018°14'326	149 m	Dusty, dry	Dry grassland: grass, *Vicia Pseudoaccacia*
W95	*L. bienne*	Croatia	N45°22'092 E016°16'175	268 m	Rocky, clay, dry, grey-brown, on limestone	Dry grassland: grass, *Achillea*, *Cichorium*, *Rubus*, *Tilia*, *Taraxacum*
W96	*L. bienne*	Croatia	N45°25'711 E015°22'231	180 m	Rocky, clay, dry, grey-brown, on limestone	Dry grassland: grass, *Daucus*, *Rubus*, *Calluna*
W97	*L. tenuifolium*	Slovenia	N45°34'913 E013°52'781	417 m	Rocky, dry, black, on limestone	Dry grassland: grass, *Pinus*

Historic cultivars were sampled from desiccated remains (Table **2.2**). From the historic seeds collection, 20 accessions were sampled from the Economic Botany Collection, Royal Botanic Garden, Kew, UK. Material sampling was under the Loans and Sampling Policy and according to a Material Supply Agreement. Additional samples of historic landraces were taken from herbarium specimens. The University of Warsaw, Botanic Garden Herbarium was approached, and 24 samples were included in this study. Furthermore, eleven samples were obtained from the Department of Plant Sciences Herbarium, University of Oxford. At the Natural History Museum, London two collections were approached; British Herbarium (seven samples) and General Herbarium (24 samples).

Table 2.4: The list of seeds of pale flax accessions

Name	Source	ID No.	Country
W042	PGR Canada	113618	Turkey
W043	PGR Canada	113630	Turkey
W044	PGR Canada	107295	Greece
W045	PGR Canada	107296	Greece
W046	PGR Canada	19021	France
W047	PGR Canada	107257	France
W048	PGR Canada	113603	Turkey
W049	PGR Canada	113606	Turkey
W050	PGR Canada	113610	Turkey
W051	PGR Canada	113621	Turkey
W052	PGR Canada	113626	Turkey
W053	PGR Canada	113629	Turkey
W054	PGR Canada	113633	Turkey
W055	PGR Canada	13636	Turkey
W056	PGR Canada	113639	Turkey
W057	PGR Canada	113642	Turkey
W065	Own collection	NA	Croatia
W066	Own collection	NA	Croatia
W067	Own collection	NA	Croatia
W068	Own collection	NA	Montenegro
W069	Own collection	NA	Albania
W070	Own collection	NA	Greece
W072	Own collection	NA	Greece
W074	Own collection	NA	Greece
W076	Own collection	NA	Greece
W077	Own collection	NA	Greece
W081	Own collection	NA	Greece
W082	Own collection	NA	Greece
W085	Own collection	NA	Bulgaria
W094	Own collection	NA	Croatia
W095	Own collection	NA	Croatia
W096	Own collection	NA	Croatia

During the collection expedition to the Balkan countries 34 samples of wild flax populations were obtained (Table 2.3). Plants were identified as *L. bienne* (16), *L. tenuifolium* (6), *L. austriacum* (4), *L. flavum* (1), *L. decumbens* (1), *L. strictum* (1)

and *L. catharicum* (1). Plants of the remaining 4 populations were not identified, but it the possibility of them being *L. bienne* could be excluded. The typical habitat of *L. bienne* is dry grasslands. It was observed that this species prefers rocky soil, but sometimes grow on sandy or clay soils. It was often spotted on hill slopes, on the edges of grassland. For the purpose of this project only *L. bienne* accessions were used. Additionally to 16 collected samples, another 16 were contributed by collaborators at Plant Genetic Resources of Canada, Agriculture and Agri-Food Canada, Saskatoon Research Centre (Table **2.4**).

2.4 DISCUSSION

The assembled plant material of domestic flax covers the areas of its cultivation in Europe and the Near East sufficiently. Most samples were collected from Turkey, France, Germany, Hungary, Czech Republic, Romania, Poland and Russia. Samples of modern cultivars cover fibre (36) and oil (22) flax, additionally there are four samples of dehiscent variety and nine landraces. Of the fibre flax accessions, only nine were identified as convariety *elongatum*, while between of the oil flax accessions, only one matched the morphological description of convariety *mediterraneum*. The sample numbers of specialized convarieties is low because they are generally very rare in seed resources around the world (Diederichsen 2007). The material acquired from the Vavilov Institute was not assessed for morphological similarity to flax convarieties. Seed material was supplemented with herbarium and museum material. The rationale for this is that specimens collected before the Green Revolution and the globalization of seed resources contain stronger phylogeographic signal at a molecular level (Lister *et al*. 2010). An excess of historic material was collected because a large proportion of samples is expected to be unusable for molecular studies due to DNA degradation.

The geographic coverage of pale flax seed resources was significantly improved during the collection expedition to the Balkans. Countries such as Croatia, Montenegro and Albania were not represented in pale flax collections. There are accessions available collected in Greece and Bulgaria, however, they are not characterized with precise geographic locations nor with data about soil and habitat. Newly collected samples from the Balkans contain geographic coordinates and detailed description of their habitat. This set combined with samples collected in Turkey (Uysal et al. 2012) represents good geographic coverage of Asia Minor and the Balkans, the two regions that separate the Near East and Central Europe. This expedition should be followed by additional collection efforts to increase the number of flax wild relative accessions in global gene banks. An increased number of samples will allow for more detailed analyses of pale flax populations and increase the resolution of a phylogeographic approach.

CHAPTER 3: IDENTIFICATION OF FLOWERING GENE ORTHOLOGS AND THEIR UTILITY FOR POPULATION STUDIES IN FLAX

3.1 INTRODUCTION

Genes controlling phenology in plants are expected to be excellent candidates for investigation of latitudinal adaptation. Daylength and consequently plant flowering strategies, change with latitude. In optimal growing conditions late determined flowering is preferable because a longer vegetative phase will ultimately increase seed production. In the northern climate however, growth indeterminacy and/or early flowering are common adaptations that allow reproduction before winter. In this way the flowering regime, by balancing resource accumulation and stress avoidance, will impact the survivability of its progeny and also crop yield. In flax, very little is known about the mechanisms of the flowering and determinacy control. The gene network responsible for flowering time is homologous between flowering plants (especially eudicots). Hence, sequences of flowering genes in *A. thaliana* and *Populus* sp. could be used to identify homologous members of the floral pathway in flax from the unannotated genome sequence. This approach would broaden our knowledge about molecular mechanism of flowering in flax, which in turn would provide a foundation for the investigation of adaptation to northerly latitudes.

In this chapter an effort has been made to broaden our understanding of flax floral genetic machinery, primarily by re-sequencing the structural homologs of core flowering time genes. The gene structures in flax, the level of their conservation and sequence similarity when compared to *A. thaliana* and poplars is discussed. Additionally, successfully sequenced loci were evaluated as molecular markers for population study.

3.2 MATERIALS AND METHODS

3.2.1 Selection of candidate genes

Among genes that are involved in flowering time control eleven were chosen according to 1) the degree of negative pleiotropism in *A. thaliana* and 2) availability of orthologous sequences in eudicot organisms from GenBank database. Genes with negative pleiotropism were not taken into account as their mutants are often under purifying selection. Loss-of-function mutations in such genes would have more than just one consequence. For example the mutation in pleiotropic gene could confer early flowering but at the same time cause lack of functionality in another important trait. Such mutations will often be selected against. Roux *et al.* (2006) proposed that non-synonymous mutations in pleiotropic genes are less likely to play role in adaptation through change in flowering time. Based on that, they ranked flowering genes into three categories: A – with no pleiotropic effects observed, B – with pleiotropic effects observed for life-history traits and C – with pleiotropic effects observed for seed production. In this way genes with A rank are more likely to carry adaptive mutations, while for the C-ranked genes the opposite is true. Availability of orthologous sequences in eudicots from GenBank database was taken into account for practical reasons: it is easier to design effective degenerate primers for target gene when multiple alignments, which they are based on, are representative for broader taxonomic group.

3.2.2 PCR survey of flowering time candidates in flax

Degenerate primers were designed to cover between 400 and 1800 bp fragment within selected floral genes. Primers were designed based on multiple alignments (Electronic Supplement **1**), which contain orthologous sequences from eudicots organisms (Table **2.1**). If substitution was present in any of organisms, the original nucleotide was replaced with degenerate nucleotide, which is complementary to both variants. Such degenerate primers increase the chance of successful amplification of targeted loci in flax, for which specific sequence is unknown. All loci were amplified by the polymerase chain reaction (PCR) with Platinum® Taq Polymerase (Invitrogen) and genomic DNA of M044 accession (Table **2.1**). PCR was carried out in TaKaRa Gradient Thermocycler with the following programme: initial denaturation in 94°C for 3 minutes, 32 cycles with sequence: 94°C for 45s, gradient temperature (50°C - 60°C) for 45s and 72°C for 75s, completed with the final

elongation step in 72°C for 5 minutes. Theoretical melting temperature of primers is between 58 and 60°C, it is recommended to test annealing temperatures lower by 10 °C for PCR with degenerate primers. Samples of PCR product along with 100bp size ladder (Fermentas) were visualized under UV light after electrophoresis in GelRed™-stained, 1.5% agarose gel. PCR products were purified using QIAquick® PCR Purification Kit (Qiagen) followed by sequencing reaction with BigDye® Terminator. Sequencing fragments were then analysed in ABi Prism® Genetic Analyzer (Applied Biosystems). Sequenced amplicons with chromatograms, in which each position was occupied by signal for only one nucleotide, were used for further analyses. Chromatograms of both strands were manually checked and assembled in Codon Code Aligner v3.7.1.2. Sequences were then used to search for similar genes within NCBI database using *BLASTN* algorithm (Altschul *et al.* 1990) in order to confirm that amplified products were orthologous to the flowering-time genes. Verified candidate amplicons were taken forward for further analyses.

3.2.3 Designing specific primers for putative flowering gene homologs

Candidate amplicons were subjected to BLAST searches against the cultivated flax genome scaffolds v1.0 and CDs v1.0 databases (Rowland & Cloutier 2012) in order to identify full sequences of genes of interest and their flanking regions. All sequences with an e value below 0 were downloaded. The search procedure in CDs v1.0 database (Rowland & Cloutier 2012) was repeated, this time by querying with sequences downloaded from the first round of BLAST search. Resulting sequences were then subject to comparison of gene structure: sequences with the same number of exons (plus/minus one) as in the original *A. thaliana* gene were kept for further analysis. This allowed for filtering out pseudogenes and assembly artefacts. Within remaining putative genes open reading frames (ORFs) were estimated using *FGENESG* tool from Softberry® package (Solovyev *et al.* 2006). After that, specific primers were designed to cover all the exons and introns in confirmed putative flowering genes. Sequences of these primers and predicted amplicon lengths are in Table **3.3**. All primer sets are characterized with melting temperatures of 60°C and hence temperature of 58°C was used for annealing step in PCR. PCR and sequencing reactions were carried out with the same reagents and equipment as described above (Section ***3.2.2***). Chromatograms were processed and resultant sequences used to

validate the presence of putative flowering gene homologs in cultivated flax genome and for further analyses.

3.2.4 Phylogenetic analyses of candidate genes in Arabidopsis thaliana, poplar and flax

Coding sequences of all the putative flowering gene homologs in cultivated flax were translated into amino acids using Translate tool from ExPASy Bioinformatics Resource Portal. Sequences of homologous proteins in *A. thaliana* and *P. nigra* were downloaded from TAIR and DDJB databases respectively. All the sequences were then aligned in *CLUSTALX* v2.1 (Larkin *et al.* 2007) and modified manually in Mesquite v2.74 where necessary. The best model of protein evolution was selected using *BIONJ* algorithm (Gascuel 1997) and Bayesian Inference Criterion (Schwarz 1978). Phylogenetic trees were estimated using Bayesian Inference in *MRBAYES* v3.2 (Ronquist *et al.* 2012) with four chains in 1.000.000 generations using best-fitted model of evolution. The final tree with posterior probabilities for each clade was obtained by summarizing sampled trees excluding 25% of initial results. Furthermore, the predicted protein sequences of flowering gene homologs in flax were compared to the models of active centres and binding pockets within respective genes in *A. thaliana* and poplars to test if crucial amino acids were conserved.

3.2.5 Assessment of the number of polymorphic sites in TFL1 homologs

A small-scale re-sequencing experiment was conducted to assess utility of candidate genes as molecular markers for addressing this project's hypothesis. Using the specific primers designed in Section ***2.3.3***, a PCR was carried out for six samples representing four distinct varieties of cultivated flax (oil, fibre, intermediate and indehiscent) and two pale flax populations from northern Balkans and Turkey. This choice of samples should give a proxy for determination of possible post-domestication gene flow events. Successful amplification of candidate genes was followed by purification and sequencing (as described in Section ***2.2.2***). Polymorphic sites were mapped and compared between the two pale flax populations and four convarieties of cultivated flax.

3.3 RESULTS

3.3.1 Selection of candidate genes

Eleven genes were chosen for the search of their putative homologs in flax. Six genes with A-rank of pleiotropism were selected on the basis of being represented in NCBI database by at least three orthologs within eudicots. Additionally, four genes with B-rank of pleiotropism and one with C-rank of pleiotropism were also chosen (Table **3.1**).

Table 3.1: The list of the flowering time genes investigated in flax and the degenerate primers used for their amplification.

Locus	Full name	Cat.	Sequences for alignment		Degenerate primers 5' to 3'
AP1	*Apetala 1*	A	AY515153.1	(*Populus deltoides*)	ATGGGRAGRGGTAGRGTTCA
			AY770395.1	(*Lotus corniculatus*)	GGTTYCTCTSCAARAGBTCAA
			NM_105581.2	(*Arabidopsis thaliana*)	
			AF466785.1	(*Arabidopsis thaliana*)	
FLC	*Flowering locus C*	A	FJ347968.1	(*Cichorium intybus*)	ATCGATATGGGAAACAGCATG
			NM_121052.2	(*Arabidopsis thaliana*)	CCAAAACCTGGTTCTCTTCTT
			AY769360.1	(*Arabidopsis thaliana*)	
TFL1	*Terminal flowering locus*	C	FM999796	(*Rosa wichurana*)	TSATGAYWGAYCCWGATGTTC
			AB383157	(*Cucumis sativus*)	CACAWAYCTRTGKRTYCCTAT
			AB369067.1	(*Populus nigra*)	
			D87519.1	(*Arabidopsis thaliana*)	
LFY	*Leafy*	A	AY640314	(*Eucalyptus grandis*)	GAGGAGCTYGAMGASATGATGA
			AY639379	(*Hevea brasiliensis*)	GTGCYACYTCCCCRGGCTC
			NM_125579	(*Arabidopsis thaliana*)	
			AF466801.1	(*Arabidopsis thaliana*)	
SVP	*Short vegetative phase*	A	FJ373210	(*Poncirus trifoliata*)	GTTCTYTGCGAYGCCGATGT
			KC297693.1	(*Populus tomentosa*)	AGRGMCTTCTCTAGCTKCTG
			NM_001161056	(*Arabidopsis thaliana*)	
			GQ177854	(*Arabidopsis thaliana*)	
FRI	*Frigida*	A	AY561834.1	(*Allium sativum*)	GCGAAGTTYGTRTTGGAWTGT
			JN047764.1	(*Populus balsamifera*)	TCCAWACACTKCRTMACTGAT
			AF228500.1	(*Arabidopsis thaliana*)	
			AY198403	(*Arabidopsis thaliana*)	
ELF3	*Early flowering 3*	B	EF185298.1	(*Pisum sativum*)	AGAGCWCCTCCYAGRAACAA
			JN047731.1	(*Populus balsamifera*)	AACTCAAAYASTTGRACWGCAA
			NM_128153	(*Arabidopsis thaliana*)	
			AT2G25930.1	(*Arabidopsis thaliana*)	
HOS1	*Osmotically responsive 1*	B	FJ844367	(*Citrus trifoliata*)	AGTGATCCKGTCRTTGCTTTC
			NM_129540	(*Arabidopsis thaliana*)	GAGACCTCACATYMTCCAGAA
			GQ177486	(*Arabidopsis thaliana*)	
PHYA	*PhotochromeA*	B	XM_002318877	(*Populus trichocarpa*)	GATATGCTSATGCGTGATGC
			NM_100828	(*Arabidopsis thaliana*)	CTTGCRCAKGCATTWACAACT
			AT1G09570.1	(*Arabidopsis thaliana*)	
TOC1	*Timing of cab1*	B	AY611028.1	(*Castanea sativa*)	GGTRATTGATGCACTKAATGC
			EU076435	(*Glycine max*)	TCATCTGTGTCRTCWGARAACA
			HQ833401.1	(*Populus tremula*)	
			AF272039.1	(*Arabidopsis thaliana*)	
			AT5G61380.1	(*Arabidopsis thaliana*)	
VRN1	*Vernalization1*	A	AY517929.1	(*Brassica rapa*)	ATTCAAGGAYGAGCTWTCGG
			AK230388.1	(*Brassica rapa*)	TGACTCGAWAGGCYGTCAC
			AT3G18990	(*Arabidopsis thaliana*)	

3.3.2 PCR survey of flowering time candidates in flax

Out of eleven degenerate primer pairs, designed to cover introns of floral genes, more than half allowed for successful amplification. The products with approximately correct size were obtained for *FLC*, *TFL1*, *LFY*, *SVP, ELF3*, *HOS1* and *TOC1* (Figure **3.1**). In case of *TFL1*, three homologs were amplified. Each product was excised from the gel using sterile blade and purified separately. Cleaned PCR products were sequenced and subsequently subjected to a BLAST search against the NCBI nucleotide collection. Only three of the amplified products were identified to match homologs of the expected target genes (Table **3.2**). These bore significant similarity to genes *TFL1*, *LFY* and *TOC1* (Electronic Supplement **2**).

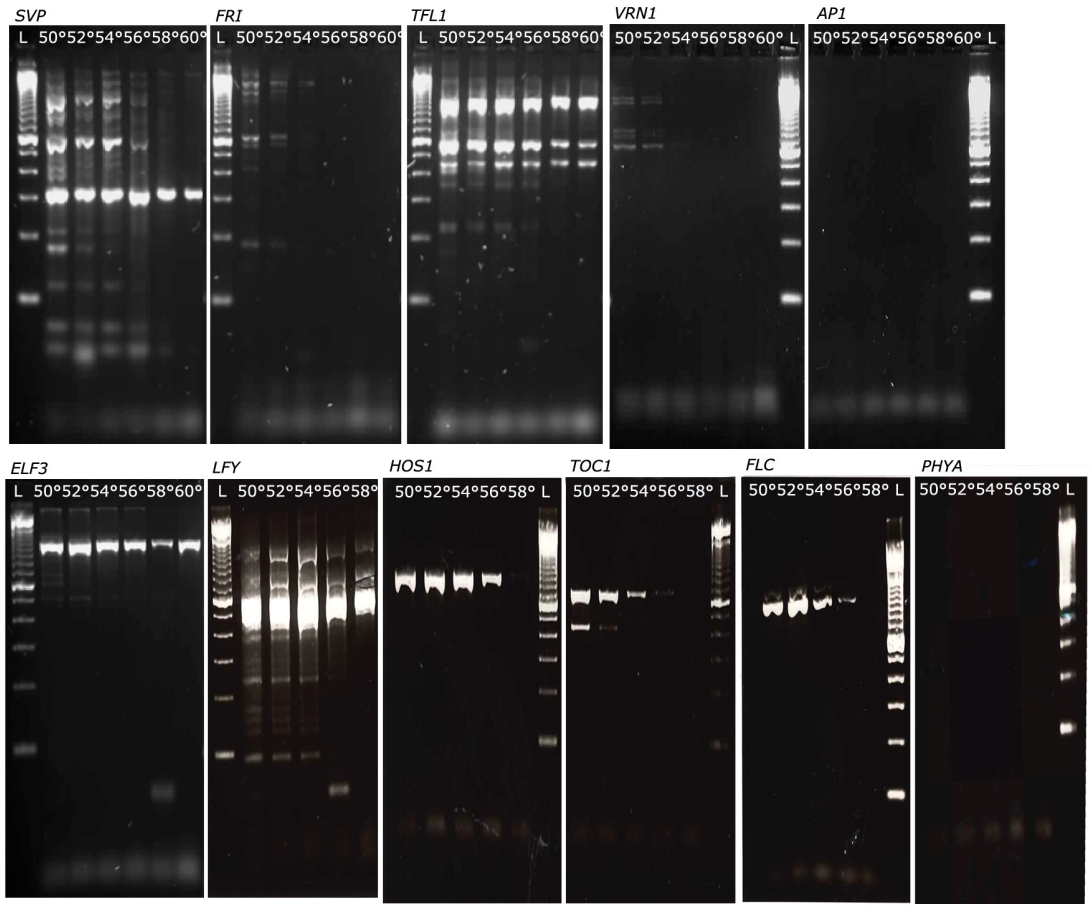


Figure 3.1: Electrophoresis gels showing amplification products of investigated flowering time genes. PCR reaction was carried out with degenerate primers in gradient of temperatures (50°C to 60°C). The product along with 100 bp ladder was visualized under UV light.

3.3.3 Designing specific primers for putative flowering gene homologs

Amplicons that matched genes *TFL1*, *LFY* and *TOC1* were used to identify full sequences of putative flowering genes in flax through BLAST search against the flax scaffold genome. These sequences matched multiple entries in the flax genome with significant e-values. Putative homologs of all three genes were downloaded (Table **3.3**). Further analysis, where initial BLAST hits were used as queries for the second round of searches, did not provide any additional results. In total, eight homologs of *TFL1*, four homologs of *TOC1* and two homologs of *LFY* were found.

Table 3.2: The list of successfully amplified and sequenced putative flowering time genes.

Locus	Expected product size [bp]	Observed product size [bp]	Clear sequence	BLAST result
AP1	1790	-	–	–
FLC	830	1100	+	–
TFL1	494	460/600/1100	+	+
LFY	985	700	+	+
SVP	1785	300	+	–
FRI	1129	-	–	–
ELF3	986	1100	–	–
HOS1	953	1000	+	–
PHYA	1102	-	–	–
TOC1	704	600	+	+
VRN1	1251	-	–	–

Coding sequences of *TFL1*, *LFY* and *TOC1* genes from *A*. *thaliana* were used to BLAST against linum.ca database (Rowland & Cloutier 2012) and to calculate sequence similarity to each putative homolog in flax. Similarities, expressed in e-values of BLAST searches are listed in Table **3.3**. *TFL1* homologs were taken at this stage for further analysis owing to the large number of multiple paralogs in flax and short length of the whole sequences. *LFY* and *TOC1* were discarded at this point (the rationale for this decision are discussed in Section ***3.4.1***).

Open reading frames and gene structures in all of the putative *TFL1* homologs in flax were estimated (Figure **3.2**). Most of them contain four exons interrupted with three introns. In flax, the first exon varies in size from 198 to 225 bp, while the final exon between 224 and 237 bp. The two inner exons have very consistent length between all the *TFL1* homologs; 62 bp in case of exon 2 and 41 bp in case of exon 3. *LuTFL5* and *LuTFL6* are the exceptions from the common *TFL1* gene structure; exon 3 and exon 4 were probably merged into one (Figure **3.2**). In general, the gene structures within flax putative homologs are very similar to these of *TFL1* and *FT* genes in *A. thaliana*.

Table 3.3: The list of identified putative flowering time gene homologs in the flax genome.

Locus	GBrowse CDS reference	E-value (A. tha CDS)	Specific primer sequence 5' to 3'	Primer positions	Length of gene
LuTFL1	Lus10004886	9e-74	GGAGAAATCGATGGGGAAAG TGGGTCGGTACTAACGCCT	+5 +1540	1496
LuTFL2	Lus10043385	4e-65	AATCGTAAAATTAACGGCGTC AACCTCGACTATGTCGTCAGA	-1 +1194	1180
LuTFL3	Lus10020600	2e-76	TGGGGCAAAATGGGGAAA ATCGATCAATCTAGCGCCG	-9 +1762	1752
LuTFL4	Lus10004884	1e-34	ATTAGGCAAAATGGGGAAAG ATCGATCAGTTTAGCGCCG	-10 +2545	2535
LuTFL5	Lus10027442	9e-30	AGTTACAACAATGGCAAGAGG TATTATTAGCGTTTTCTGGCG	-10 +1268	1263
LuTFL6	Lus10005753	2e-26	AGTTACAAAAATGGCAAGAGGA TGGCGGCATATTAGCGTT	-10 +1196	1185
LuTFL7	Lus10021372	2e-64	AAAAGATGGCAGCAGGTGC TAGATCGATGTCATCGACGC	-5 +1897	1877
LuTFL8 (*LuFT*)	Lus10004452	2e-50	GTTATCGAAAATGCCAAGGG GAGGATATCATCATCGCCGT	-10 +1841	1832
LuTOC1	Lus10021629	3e-105	N/A	N/A	N/A
LuTOC2	Lus10019077	1e-98	N/A	N/A	N/A
LuTOC3	Lus10015720	5e-96	N/A	N/A	N/A
LuTOC4	Lus10000566	6e-95	N/A	N/A	N/A
LuLFY1	Lus10016732	4e-114	N/A	N/A	N/A
LuLFY2	Lus10022427	1e-107	N/A	N/A	N/A

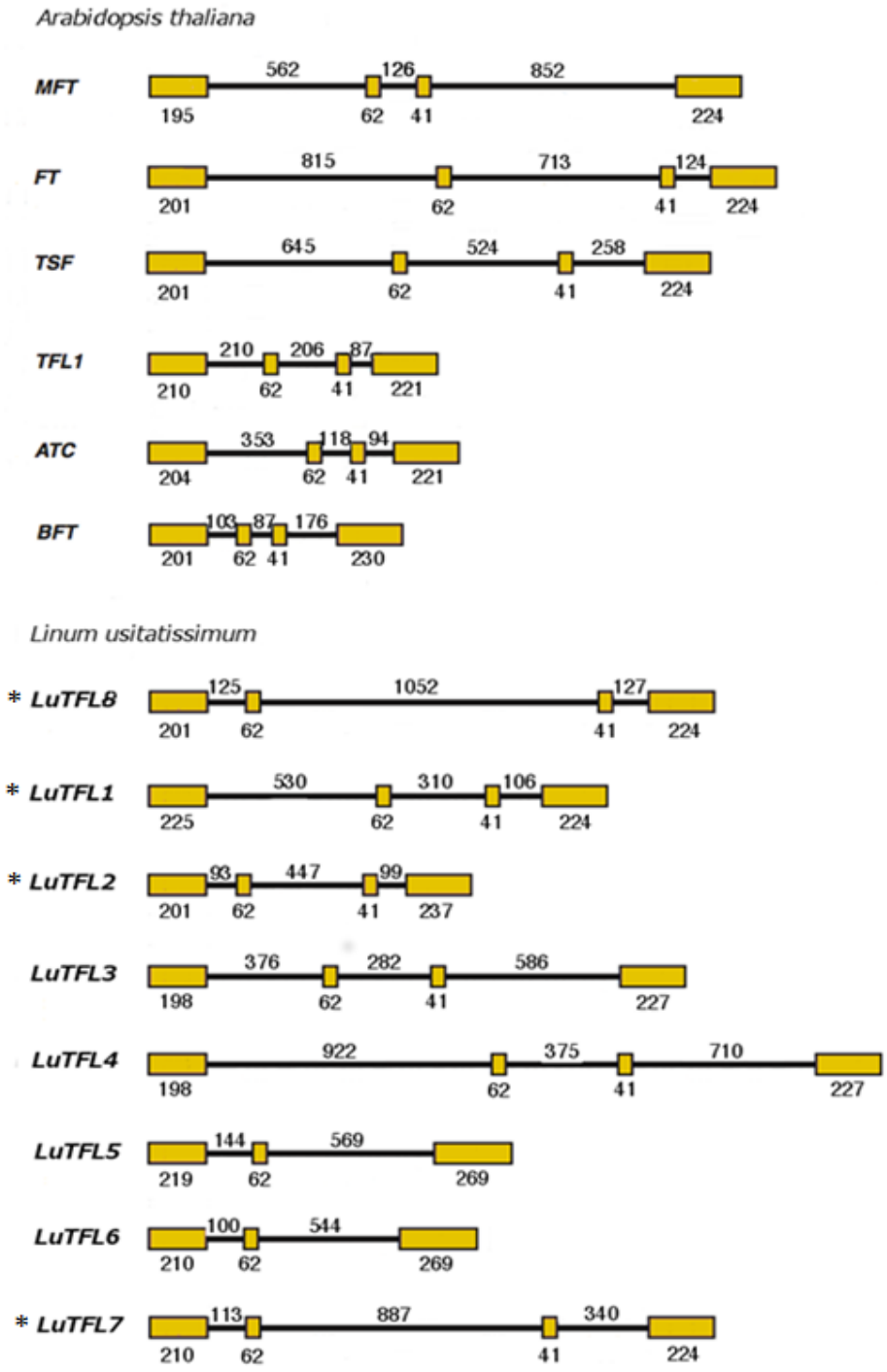


Figure 3.2: Structure of PEBP family genes in *Arabidopsis thaliana* and putative homologs in flax. The numbers above genes denote lengths of intronic and exonic regions. The flax sequences marked with asterisks are a result of successful sequencing and ORF establishment, the remaining ones are predicted genes from linum.ca database (Rowland & Cloutier 2012).

The sequences of flanking exons were used to design specific primers that allow for the whole gene amplification. Such primers were designed for all eight putative homologs (Table **3.2**). After amplification with specific primers seven out of eight putative homologs of *TFL1* yielded positive results (Figure **3.3**). It was problematic to design primers that were specific enough to amplify just one locus due to very high similarity of sequences in pairs *LuTFL3*, *LuTFL4* and *LuTFL5*, *LuTFL6*. Chromatograms of the successful amplicons were used to validate sequences of putative *TFL1* homologs in flax. The sequences of pair *LuTFL3*, *LuTFL4* were identical while *LuTFL5*, *LuTFL6* were ambiguous and unreadable (Electronic Supplement **2**).

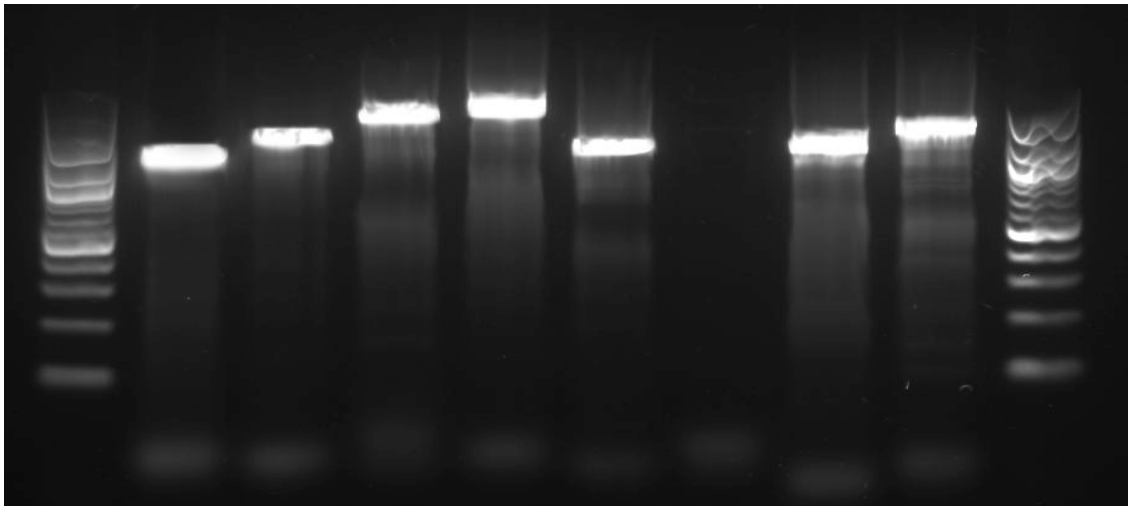


Figure 3.3: Electrophoresis gel showing amplification products of eight *LuTFL* paralogs in flax. PCR reaction was carried out with specific primers for *LuTFL1*, *LuTFL2*, *LuTFL3*, *LuTFL4*, *LuTFL5*, *LuTFL6*, *LuTFL7* and *LuTFL8* subsequently. The product along with 100 bp ladder was visualized under UV light.

3.3.4 Phylogenetic analyses of candidate genes in Arabidopsis thaliana, poplar and flax

Tentative functional assignations were carried out through a phylogenetic analysis comparing the flax *LuTFL* homologs to characterized members of the wider PEBP family in the model organisms of *Arabidopsis* and *Populus*. Since not all of the *LuTFL* chromatograms were clear enough for sequences to be used in phylogenetic analyses, the CDs predicted by linum.ca (Rowland & Cloutier 2012) were used instead. These sequences together with PEBPs from *A. thaliana* and *P. nigra* were translated into amino acids and aligned (Table **3.4**). BioNJ algorithm suggested JTT+G model of evolution as the best fitting to the data. Bayesian inference,

phylogenetic analysis was carried out and resultant tree revealed relationship between flax putative *TFL1* homologs and genes from PEBP family in *A. thaliana* and *P. nigra* (Figure **3.4**). None of the flax sequences showed similarity to ancestral *MFT* genes. Well-supported clades separated *FT*-related genes from *TFL* homologs. Only one of the putative homologs in flax – *LuTFL8* was assigned to *FT* clade, while remaining seven grouped with *TFL*-like genes. Within these, both of the 3-exonic putative genes in flax (*LuTFL5* and *LuTFL6*) are clustered with *BFT*, while *LuTFL7* is most similar to *TFL1*. The remaining putative *TFL* homologs in flax (*LuTFL1*, *LuTFL2*, *LuTFL3* and *LuTFL4*) form a clade with *CEN* in *A. thaliana* and *P. nigra*, but with a branching order conflicting with the species relationships so are considered here to be paraphyletic to CEN with high posterior probability. It is likely that this group of flax *TFL*s have functions similar to that of *CENTRODADIALIS* and/or *TFL1*.

Predicted protein sequences for *LuFT* and *LuTFL* homologs showed high similarity to *A. thaliana* PEBP genes on phylogenetic tree. A closer look into actual sequences revealed that all the amino acids forming binding pocket and activity-specific residues in *A. thaliana* were conserved in flax.

Table 3.4: The list of PEBP gene family homologs used for phylogenetic analyses.

Locus name	Organism	Database	Accession No.	Homolog relative to A. thaliana
MFT	*A. thaliana*	TAIR	NM_101672	NA
BFT	*A. thaliana*	TAIR	NM_125597	NA
FT	*A. thaliana*	TAIR	NM_101672	NA
TSF	*A. thaliana*	TAIR	NM_118156	NA
TFL1	*A. thaliana*	TAIR	NM_120465	NA
CEN	*A. thaliana*	TAIR	NM_128315	NA
PnFTL1	*P. nigra*	DDBJ	AB161110	*BFT*
PnTFL1	*P. nigra*	DDBJ	AB181183	*CEN*
PnFTL3	*P. nigra*	DDBJ	AB181185	*CEN*
PnFTL4	*P. nigra*	DDBJ	AB181241	*MFT*
PnFT1	*P. nigra*	DDBJ	AB106111	*FT*/*TSF*
PnFT2	*P. nigra*	DDBJ	AB109804	*FT*/*TSF*
PnFT3	*P. nigra*	DDBJ	AB110612	*FT*/*TSF*
PnFT4	*P. nigra*	DDBJ	AB369074	*FT*/*TSF*
PnFTL9	*P. nigra*	DDBJ	AB369072	*FT*/*TSF*

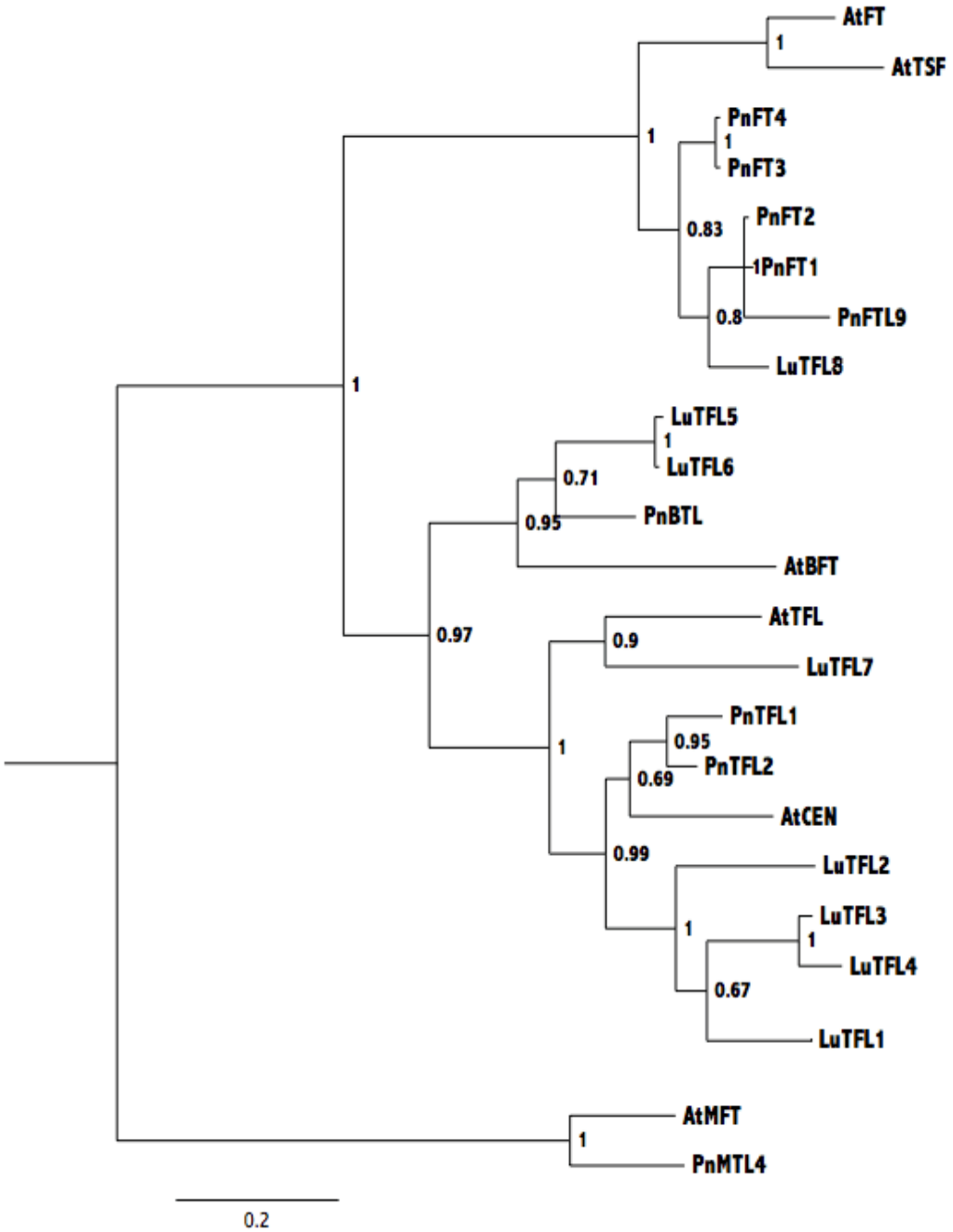


Figure 3.4: Bayesian inference phylogenetic tree representing relationship of PEBP family genes in *Linum usitatissimum*, *Arabidopsis thaliana* and *Populus nigra*. Values by the internal nodes represent clade posterior probability. Gene prefixes indicate the organism in which locus was identified (At – *A. thaliana*, Lu – *L. usitatissimum*, Pn – *P. nigra*).

3.3.5 Assessment of the number of polymorphic sites in TFL1 homologs

Evaluation of the number of polymorphic sites within the flax *TFL* loci was possible after preliminary re-sequencing in six different accessions of flax. These accessions were representing the four distinct convarieties of cultivated flax and the two pale flax populations – from northern Balkans and Turkey. The aim of this experiment was to identify polymorphic sites that allow distinguishing northern and southern pale flax and at the same time segregate in different flax varieties. In *LuTFL1* 22 polymorphic sites were discovered and they allowed distinguishing the northern and southern pale flax. Moreover, based on this marker, the northern population shows high similarity to cultivated flax. One of the polymorphic sites accommodated a non-synonymous mutation, which turned a tryptophan-coding triplet into arginine in exon 3. Remaining 21 SNPs were present in introns. Together they allowed for differentiation of all six samples tested.

Table 3.5: Summary of the polymorphism search within seven *LuTFL* homologs in flax.

Locus	No. of polymorphic sites	No. of polymorphic sites within pale flax	Ancestry
LuTFL1	22	19	Mixed
LuTFL2	21	1	Northern
LuTFL3	21	3	Southern
LuTFL4	N/A	N/A	N/A
LuTFL5	N/A	N/A	N/A
LuTFL6	N/A	N/A	N/A
LuTFL7	31	0	Uniform

Ancestry: the origin of cultivated flax polymorphism. Southern – mutations in southern pale flax exist in cultivated flax, northern - mutations in northern pale flax exist in cultivated flax, mixed – mutations from both pale flax samples exist in cultivated flax and uniform – there are no unique mutations within pale flax.

In *LuTFL2* 21 polymorphic sites were discovered. One of them was identified as microsatellite fragment with variable number of 'AT' repeats. In this locus haplotypes of cultivated flax are more similar to northern than to southern pale flax. *LuTFL3* accommodated 21 polymorphic sites. Here as well one microsatellite region

was found. There are only three mutations between the two pale flax samples. One of these mutations places southern pale flax closer to cultivated flax, while another indicates closer similarity of northern sample to cultivated flax. In *LuTFL4* it is difficult to calculate the number of polymorphic sites; sequences were not homologous in the whole length. Pale flax samples were aligned only to the third part of the sequence rendering the alignment unreadable. Furthermore, sequences of *LuTFL5* and *LuTFL6* were so similar that it was difficult to distinguish between them and hence they were not analysed. In *LuTFL7* 31 polymorphic sites were found and they allowed to distinguish two different haplotypes in cultivated flax, however, sequences for both northern and southern pale flax were identical (Electronic Supplement **2**).

3.4 DISCUSSION

The results show that the method of homolog mining used in this chapter for flax is not efficient but still successful. This approach is the only option in absence of genomic data. Thus, the utility of chosen target genes as molecular markers for population genetic studies is evaluated further and the structure of the few successfully resequenced putative flowering homologs is analysed and compared to the known genes in *A. thaliana* to infer about possible functional homology.

3.4.1 Evaluation of gene mining experiment

Overall, the degenerate primer re-sequencing approach applied in this study turned out to be moderately successful in mining for flowering genes in the un-annotated flax genome. Out of eleven genes targeted by PCR systems, only three turned out to cover the targeted putative homologs and these are: *TFL1*, *LFY* and *TOC1*. Amplification of *AP1* was unsuccessful, but it is hard to speculate that there is no homolog of this highly conserved gene from *MADS* family in flax. It is believed that main floral meristem identity genes are common in all angiosperms (Bowman 1997) and hence detection of *AP1* failed using degenerate primer approach. By contrast, difficulties in amplification of the *FLC* gene could possibly reflect absence of this gene in the flax genome. Despite attempts, this gene has not been isolated in related poplars (Leseberg *et al.* 2006). Although *FRI* gene is redundant without *FLC* it has been reported in *P. balsamifera*; in the same study the presence of *PHYA* in the poplar genome has been confirmed (Keller *et al.* 2011a). These two genes are likely to be present in the flax genome as well, however, remain unidentified. Four genes (*ELF3*, *HOS1*, *VRN1* and *SVP*) are known to fulfil important roles in flowering time pathway, however, they are not described as highly conserved. Hence, it is difficult to infer if failure in identifying their homologs in flax is a result of limited efficiency of degenerate primer resequencing approach or they simply do not exist in the flax genome.

In the presented study, three putative homologs of flowering time genes were identified in flax. *LFY* and *TFL1* were already sequenced during earlier studies on flax earliness (De Decker 2007). The approach presented here enabled additionally to identify a putative homolog of *TOC1* gene. The degenerate PCR approach to mine for genes in flax was necessary in this study because at the time of the work there

was no available database of the flax genome. The degenerate primer approach allowed the identification of sequences that could then be used to subsequently search the flax genome database as it became available. The targeted putative genes shared similarity to exons of re-sequenced amplicons but would be more difficult to find when database was queried with *A. thaliana* sequences. No significant hits were found when the flax scaffold genome was queried with full genomic and coding sequence of *A. thaliana*.

TFL1, *LFY* and *TOC1* represent very important genes of the photoperiod-dependent pathway that regulates flowering time. This list could potentially be expanded by adding *CO* and *GI* genes, which mediates between *TOC1* and *FT*/*TFL1*. In this study however, these genes were omitted as they are thought not to be under selection for early flowering (Roux *et al*. 2006). Genes identified in flax do not represent vernalization-dependent pathway. *FLC* and *FRI* were not found in flax. The optimization of PCR conditions and degenerate primer concentrations might have improved the chances of amplifying these genes. Successful sequencing of *FLC* and *FRI* in flax would be very desirable, however as discussed above, it is likely that homologs of these genes do not exist in flax genome.

All three putative flowering genes that successfully retrieved from flax in this study (*TFL*, *LFY* and *TOC*) led to the identification of multiple homologs through a BLAST search against the flax genome scaffolds. In case of *LFY* and *TOC1*, two and four significantly similar loci were identified respectively. Neither of these genes have got paralogs in the *A. thaliana* genome. The putative *TFL1* homologs are interesting in that eight of them were identified, suggesting a paralogous expansion of this gene either in flax or the lineage leading to flax subsequent to the split leading to *Populus*. The number of *FT*/*TFL1* homologs in flax is larger than number of PEBP family genes in *A. thaliana* (for more details see Section ***1.4.3***) but lower than number of *FT*/*TFL* genes in *P. nigra* (Igasaki et al. 2008). Out of eight only seven *TFL* putative homologs in flax were amplified using specific primers (Figure **3.3**). Sequences in pairs *LuTFL3*, *LuTFL4* and *LuTFL5*, *LuTFL6* were ambiguous and very similar and hence it is highly possible that their discrimination might have been an error introduced during the flax genome assembly. In summary, it is confirmed that loci *LuTFL1*, *LuTFL2*, *LuTFL7* and *LuTFL8* exist in the genome of cultivated flax, whereas loci *LuTFL3/4*, *LuTFL5/6* require more in-depth investigation.

3.4.2 The choice of molecular markers for population studies

The choice of putative genes for resequencing on population scale (experiment carried out in **Chapter 3**) was based on 1) difficulty of amplification, 2) literature review, 3) number of polymorphic sites per sequence and 4) preliminary analyses of relationship within six representative flax accessions. Based on the first criterion *TOC1* was discarded. This gene is longer than 3 kb, which renders it difficult to sequence especially in presence of four putative homologs. The full *LuLFY1* gene is about 4160 bp long and there is a risk of confusion with its paralog *LuLFY2* if partial regions are amplified and sequenced. In case of both these genes, there would be the technical issues associated with gene amplification and to that, neither of these genes are commonly used in population genetics. Both *LFY* and *TOC1* has been resequenced in studies of latitudinal adaptation in poplars (Keller *et al.* 2011a), however, to the best of the author's knowledge they have not previously been used in the context of domestication.

TFL1 homologs in flax are good candidates for markers in population genetics approach. Firstly, the *FT*/*TFL* putative genes in flax are much shorter than *LFY* and *TOC1*. Their size range from 1171 to 2535 bp and thus they are short enough for single amplicons. In case of the shorter homologs the entire gene can be sequenced in a single reaction per DNA strand. Homologs of *FT*/*TFL* gene family were used multiple times in studying domestication and post-domestication adaptation to local environments (Blackman *et al.* 2010; Comadran *et al.* 2012; Pin *et al.* 2010; Tian *et al.* 2010). Based on the above arguments the potential of *FT*/*TFL* homologs in studying flax phylogeography and adaptation to northerly latitudes is great.

The selection of markers used in the next chapter for population genetics experiment is based on the two criteria: number of polymorphic sites per sequence and preliminary analyses of relationship within six representative flax accessions. The number of polymorphic sites in *TFL1* putative homologs span between 19 and 27. In *LuTFL1* there are 22 SNPs, while in *LuTFL2* - 19. Moreover, the patterns of relationship in the preliminary data for these two genes are very promising. In the *LuTFL1* northern group of pale flax was more closely related to cultivars than southern population, whereas in *LuTFL2* polymorphism differentiated all six samples sequenced. It was not possible to calculate number of polymorphic sites for putative

loci *LuTFL3* and *LuTFL4*. Due to their high similarity, it was difficult to amplify these putative genes without risk of contamination from one another. Similar issue was observed for *LuTFL 5* and *LuTFL6*, moreover, these putative genes did not maintain gene structure with four exons. This might indicate that these loci accommodate non-functional genes and therefore could be less useful for locating signature of adaptation to the north. *LuTFL7* with 27 polymorphic sites did not allow for discrimination between northern and southern populations of pale flax. Finally, the reasons for discarding *LuTFL8*, which is identified as *FT* homolog are described below. In summary, the *LuTFL1* and *LuTFL2* putative genes are the most informative given the questions posed in this thesis. They accommodate good number of SNPs and more importantly – mutations that allow linking north populations of pale flax to cultivated flax.

Any gene of floral genetic network could alter flowering time in eudicots, however, Roux and collaborators (2006) note that not all of them have adaptive potential because of the pleiotropic effects they exhibit. This often means that cost of mutations in these genes would be too high when compared to benefits of short-day adaptation. Furthermore, loss-of-function mutations, which could be associated with simple point mutation, are more common in rapid adaptation of crop plants than gain-of-function mutations. Therefore, in order to be early flowering it is in all likelihood easier to knockout an inhibitor of flowering than to enhance or reduce the flowering promoter. Hence, *TFL1* putative homologs rather than *FT*'s were chosen for further investigation.

3.4.3 The functional assignation of FT/TFL1 gene homologs in flax

The phylogenetic and structural analyses of *FT*/*TFL1* homologs in flax allowed for determination of relationship to specific genes in *A. thaliana* and poplars. Based on phylogenetic analyses *LuTFL8* is very closely related to *FT* in *A. thaliana* (Figure **3.4**). Additionally, the length of the exons in *LuTFL8* matched exactly the length in the *FT* gene of *A. thaliana* (Figure **3.2**). Finally, the amino acids of binding pocket and active sites in *FT* protein are conserved in *LuTFL8* (Figure **3.5**). For these three reasons *LuTFL8* is expected to be structural and functional homolog of *FT* and therefore will be referred to as *LuFT*. In the case of the *LuTFL7* gene, the phylogenetic similarity pointed to *TFL1* as the most likely homolog. This is

supported by gene structure analyses; the only difference between *LuTFL7* and *TFL1* is an additional coding triplet in exon number four of the former. Finally, *LuTFL1*, *2*, *3* and *4* form a sister clade to *ATC* gene of *A*. *thaliana*. In these genes exon 2 and 3 are the same size as in all the PEBP genes in *A*. *thaliana*. On the other hand, exons 1 and 4 are variable, the differences from *A*. *thaliana ATC* however, do not exceed five amino acids. Finally, both *LuTFL1* and *LuTFL2* contain all active- and binding-site amino acids of *ATC* gene in *A*. *thaliana*. Remaining two homologs in flax: *LuTFL5* and *LuTFL6* were both grouped together with *BFT*, however, their exon structure is different from the latter in that their exon 3 and 4 were probably merged together. All of the studied genes in flax were paired with respective genes in *A*. *thaliana*.

The function of *FT*/*TFL* gene homologs in flax could be speculated based on the known roles of these genes in *A*. *thaliana*. For example *LuFT* could play the same role as *FT* in promoting the flowering. By contrast, *TFL1* and *ATC* homologs probably play a role in delaying the flowering. Interestingly, *TFL1* was characterized with category C by Roux and collaborators (2006), which mean that it exhibits pleiotropic effects. The nature of *TFL1* pleiotropism however, is very interesting for this study. In *A. thaliana* this gene regulates time of flowering but also mode of growth – determinate vs. indeterminate, plant height and inflorescence architecture (Bradley *et al.* 1997; Prusinkiewicz *et al.* 2007). The latter two traits are very important in distinguishing oil and fibre varieties and therefore *TFL*-like gene have huge potential in helping to answer the main question posed in this thesis: has the adaptation to northerly latitudes had an impact on evolution of specialized fibre varieties.

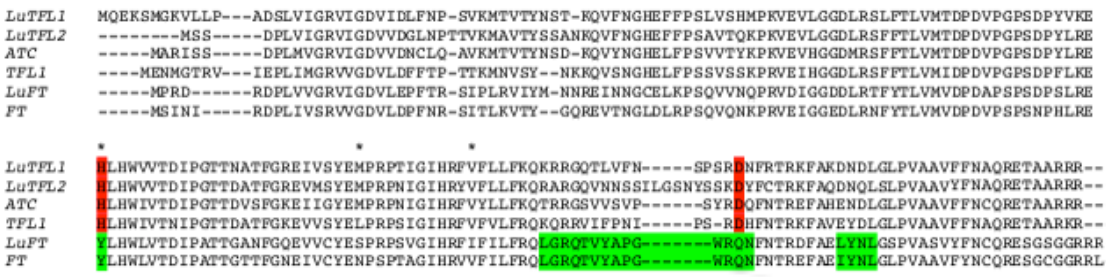


Figure 3.5: Conservation of amino acids in active sites of *FT*, *TFL1* and *ATC* genes and their putative homologs in flax. Asterisks indicate the potential binding pocket amino acids (Yoo *et al.* 2010), in red the residues that are typical for proteins with *TFL1* activity, in green – with *FT* activity (Igasaki *et al.* 2008).

All lines of reasoning point at *LuTFL1* and *LuTFL2* as the best candidates for further population genetics study in cultivated and pale flax. In *LuTFL1* we can see close relationship between northern populations of pale flax and flax cultivars that might indicate post-domestication gene flow between the two species. On the other hand, differences in *LuTFL2* sequences led to discrimination of the two pale flax populations and all of the major convarieties in cultivated flax. In the next chapter an effort to re-sequence these two putative genes in sample of 144 flaxes is described.

CHAPTER 4: PHYLOGEOGRAPHIC ANALYSES OF THE *LUTFL1* AND *LUTFL2* PUTATIVE GENES IN FLAX

4.1 INTRODUCTION

Flowering strategies of crops change with climatic conditions. Diversity in this trait is an effect of genetic variation in underlying floral genes. Alleles of these genes conferring flowering strategy, which is better adapted to local climatic conditions are preferred, and hence under positive selection (Nielsen 2005). This is reflected by increased allele frequency in the area in which specific climatic conditions apply. Therefore, adaptive alleles often are associated with geographic locations. Genotyping experiments carried out for the large set of populations allow reconstruction of the geographic associations of alleles and hence investigation of their adaptive properties. In flax, there have been no reports of genomic regions under selection, which might have contributed towards flax adaptation to different latitudes. The only success in latitudinal structuring of flax population was within flaxes in North America (Soto-Cerda et al. 2012) however, this study was based on neutral markers and therefore no inference on adaptive properties of analysed loci can be made. Additionally, winter and summer varieties are genetically distinguishable (Fu 2012) yet again this study was based on neutral markers.

In this chapter, two loci *LuTFL1* and *LuTFL2,* which encompass structural homologs of flowering genes from the PEBP family, were chosen for population resequencing in both cultivated and pale flax. Alleles of these genes are expected to play a role in changing flax flowering strategy. Both loci contain polymorphisms that allow the detection of relationships between northern pale flax and cultivated flax. In this chapter, haplotype networks of these loci were investigated and neutrality of their evolution tested. Additionally, a phylogeographic analysis was carried out. Molecular variation and allele frequencies of *LuTFL1* and *LuTFL2* genes on spatial scale might be the key to investigating flax adaptation to Central and Northern Europe.

4.2 MATERIALS AND METHODS

4.2.1 Flax accessions and DNA isolation

Within collected material, which is described in Chapter 2, a total of 148 samples were used in this study. This set includes 58 cultivars, 18 landraces, 38 historic landraces of cultivated flax, 32 pale flax samples and two *L. decumbens* accessions. The complete list can be found in Supplement **1**, detailed information about each accession can be found in Tables **2.1**, **2.2** and **2.4**. The total DNA was isolated using modified DNEasy® Plant Mini Kit (Qiagen) protocol. For modern samples, three and ten seeds were used for DNA extraction in cultivated and pale flax respectively. In case of herbarium specimens, 20 mg of mixed plant tissue was ground. Ground samples were mixed with 660 µl of 2% CTAB solution and 340 µl of MilliQ water followed by incubation in water bath at 65°C for 90 minutes. After centrifugation supernatant was transferred to 800 µl of chloroform:isoamyl alcohol (24:1) and mixed gently. This mix was centrifuged at high speed and supernatant was moved to tube with 2 volumes of AP3/E buffer followed by standard protocol. DNA concentration was measured with Quant-it™ dsDNA Broad-Range Assay kit (Invitrogen) in Qubit® fluorometer (Invitrogen).

4.2.2 Resequencing and sequence analyses

Two markers in putative flowering genes were developed in flax (for details see Section ***3.4.2***). Marker primers were designed to include all the introns of *LuTFL1* and *LuTFL2* loci. The optimization of annealing temperatures was carried out using gradient PCR with temperatures between 50 and 60°C. Touchdown approach for annealing temperatures was used to improve primer specificity in historic samples. Primer sequences and their optimal annealing temperatures are listed in Table **4.1**. *LuTFL1* and *LuTFL2* loci were amplified in PCR reaction with Platinum® Taq Polymerase (Invitrogen). For modern material the *LuTFL1* amplification steps are: 10 cycles with 94°C (45s), 59°C (45s) decreased by 0,1 with each step and 72°C (75s) followed by 25 cycles with 94°C (30s), 58°C (40s) and 72°C (75s) and final elongation step. Similarly, cycling steps for *LuTFL2* are: 10 cycles with 94°C (45s), 55°C (45s) decreasing by 0,1 with each step and 72°C (60s) followed by 25 cycles with 94°C (30s), 54°C (40s) and 72°C (60s) and final elongation step. For historic material, in which amplifiable fragments are shorter due to DNA degradation, internal primers had to be used. For these samples the amplification steps are

identical to above with change in elongation step from 60s to 30s. Sample of PCR product was visualized under UV light after electrophoresis in 1.5% agarose gel stained with GelRed™. Remaining PCR product was purified using QIAquick® PCR Purification Kit (Qiagen) and followed by sequencing reaction with BigDye® Terminator. Sequencing fragments were then analysed in ABi Prism® Genetic Analyzer (Applied Biosystems). Alternatively, samples were sequenced by outsourcing company (GATC).

Table 4.1: *LuTFL1* and *LuTFL2* primer sequences and optimum annealing temperature for PCR.

Locus	Primers	Sequence (5' – 3')	Annealing temperature
LuTFL1	Forward	TTACAACTCCACCAAGCAAGTC	59°C
	Reverse	TGTCTCGCGCTGAGCATT	
	Internal forward	TCTATAACAACATGAATTTGGTAT	
	Internal reverse	CCAGTGTAAATGCTCTTTCAAG	
LuTFL2	Forward	TCGTCAGATCCTCTAGTGATA	55°C
	Reverse	TGATTGATGTTATGTGTGTATGG	
	Internal forward	GAAGCTCTCATAAGAGCTAAC	
	Internal reverse	TTTGGGTTGGCACTCTTACAA	

Chromatograms of both strands of *LuTFL1* and *LuTFL2* were manually assessed and contigs were assembled in *CODON CODE ALIGNER* v3.7.1.2. Chromatograms for heterozygotes that were ambiguous and unclear (effect enhanced by indels associated with short tandem repeats) within *LuTFL1* and *LuTFL2* sequences were rejected from further analyses. Remaining sequences were then aligned with *CLUSTALX* v2.1 (Larkin *et al.* 2007) and gaps were re-aligned manually in Mesquite v2.74 where necessary. Number of polymorphic sites in analysed dataset was calculated manually, nucleotide diversity was estimated using R package Pegas. Additionally, recombination signatures were assessed within haplotypes of *LuTFL1* and *LuTFL2* separately.

4.2.3 Phylogeography of flax

Haplotype networks for *LuTFL1* and *LuTFL2* separately were created using Uncorrected_P character transformation and RootedEqualAngle splits transformation

in *SPLITSTREE4* (Huson & Bryant 2006). Size of the network nodes was made proportional to the number of samples that share a particular haplotype. Almost half of pale and cultivated flax samples were characterised with precise geographic location. For remaining samples of cultivated flax only country of origin was known. In these cases averaged latitudes (centroids) were established based on the locations provided by the Geographic Names Server (National Geospatial-Intelligence Agency 2012). Based on latitude data the nodes of haplotype network were divided into sectors and marked with different shades: from light which denote southern latitudes to dark denoting northerly latitudes. Furthermore, relationship between latitude and *LuTFL1* and *LuTFL2* haplotypes separately was modelled using logistic regression in *R* programme and plotted using *POPBIO* package. Increase of haplotype frequency per grade of latitude was calculated using logit transformation.

For the majority of wild flax samples (28/32), precise geographical data are available and these were used in testing spatial autocorrelation against *LuTFL1* and *LuTFL2* haplotype data. Classic Mantel's test of matrix congruence (Mantel 1967) was carried out together with custom-made permutation test. This test written in *R* by the author of this thesis computes sum of average pairwise distances within haplogroups (s-score). Furthermore, it compares the s-score for observed data to the s-scores for permuted data, in which haplotypes are randomly dispersed on spatial scale (source code attached as Electronic Supplement **3**). When s-score for observed data is lower than 5% of s-scores from permuted data then it shows non-random distribution of haplotypes on spatial scale.

4.2.4 Linkage disequilibrium and signature of selection

Linkage disequilibrium (LDE) was measured within and between all the SNPs of the two investigated loci to establish if *LuTFL1* and *LuTFL2* are inherited independently. For this purpose, *LDHEATMAP* package was employed in *R* to enable calculation of Lewontin's D' (Lewontin 1964) and squared allelic correlation r^2 (Pritchard & Przeworski 2001) as the measure of LDE. Neutrality tests were also carried out to test if these loci were inherited neutrally or were under selection. Tajima's D (Tajima 1989), Fu and Li's D, Fu and Li's F (Fu & Li 1993) statistics were calculated in *INTRAPOP* (Achaz 2013) and R2 statistic (Ramos-Onsins & Rozas 2002) was calculated in *R* using *PEGAS* package.

A custom forward model was scripted in *PERL* language (source code attached as Electronic Supplement **4**); its role is to establish the selection rates during adaptation of cultivated flax to the northerly latitudes. It was used to simulate cultivated flax adaptation under three scenarios: 1) presence of advantageous allele within cultivated flax gene pool in the area of domestication, 2) acquisition of advantageous allele from pale flax through introgression with Balkan populations and 3) combination of both. In these experiments 1000 plants were grown for 12 000 generations reflecting 12 000 years of cultivated flax evolution. After 3 000 generations - time required for cultivated flax to reach Europe (as reviewed in Price (2000)) and corresponding to flax archaeological findings in Near East (Hillman 1975) and Greece (Valamoti 2011) selection was applied. The strength of selection determined the number of individuals without the adaptive allele that were removed from population in each generation. This parameter is set by users, the general idea behind that is to explore how different selection coefficients influence dynamics of adaptation. Simulations with wild gene flow were repeated for rates of hybridization spanning from 0.1% to 5%, the latter being the upper limit of crosspollination observed in flax field experiments (Gurbuz 1999) and congruent with migration rates estimated in wild barley (Hubner *et al.* 2012). The advantageous allele was transferred in generations: 3000 – 5000 reflecting the period of time needed to arrive to Central Europe and in which flax have not yet escaped beyond distribution of wild flax, congruent with estimates of agriculture spread (Price 2000) and flax archaeological findings in Central Europe (Maier & Schlichtherle 2011). The final frequency of advantageous allele was measured against different rates of selection in 1000 replicates. This model was compared to the observed allele frequencies. The number of replications in which advantageous allele reached the same frequency as in observed data was noted. This number was used to score probability of observed data given the different selection coefficients.

4.3 RESULTS

4.3.1 Sequence analysis

The *LuTFL1* and *LuTFL2* introns are highly polymorphic markers within both cultivated and pale flax. Most of the samples were homozygotes. However, 12 and 14 samples were discarded in *LuTFL1* and *LuTFL2* datasets respectively because of heterozygosity (all chromatograms attached as Electronic Supplements **5** and **6**). Chromatograms of these sequences were unclear and ambiguous due to sequence shifts that characterise heterozygous state in short tandem repeat regions. *L. decumbens* sequences were also discarded due to their high divergence from *L. usitatissimum* and *L. bienne* and consequent problems in aligning them to the core data. Remaining samples were accurately aligned in *CLUSTALX*, only minor changes were introduced manually in Mesquite in the gaps resulted from indels (alignments attached as Electronic Supplement **7** and **8**). Alignments of both loci were analysed for number of polymorphic sites and nucleotide diversity. There are 50 SNPs in *LuTFL1* and it is characterized by nucleotide diversity of 0.0066, while *LuTFL2* accommodate 22 polymorphic sites with diversity of 0.0113. Finally, the analysis of signature of recombination was carried out. However, this led to conclusion that there are no linkage blocks within both loci and therefore no recent recombination events can be inferred.

4.3.2 Phylogeography of flax

Haplotype networks based on *LuTFL1* and *LuTFL2* data (Figure **4.1a** and Figure **4.2a** respectively) show different patterns of diversification in cultivated and pale flax accessions. More importantly they show different ways of how pale and cultivated flax are related to one another. The haplotype information and geographic origin of each sample is detailed in Supplement **1**.

Table 4.2: Summary of sequencing data analyses.

Locus	Length of marker	Dataset	No. SNPs	Nucleotide diversity	No. samples used
LuTFL1	1306 bp	All	50	0.0066	134
		Cultivated	44	0.0047	103
		Pale	36	0.008	31
LuTFL2	1083 bp	All	22	0.0113	132
		Cultivated	20	0.009	101
		Pale	14	0.0023	31

In the *LuTFL1* based network there are ten different haplotypes associated with pale flax, where four of them represent more than one flax accession. Haplotype I groups together populations from southern and eastern Turkey, haplotype II only these originated from southern Greece. Moving further north, samples collected from Bosphorus region and middle Balkans are included mostly in haplotype III while northern-most populations form haplotype VIII. The remaining six haplotypes consist of single accessions of pale flax (IV, V, VI, VII, IX, X). The *LuTFL1* haplotypes within pale flax have a strong geographic structure as measured by spatial autocorrelation; the null hypothesis of a random geographic distribution was rejected by Mantel's test and a custom made permutation test (Table **4.2**).

Table 4.3: P-values of spatial autocorrelation tests for pale flax samples.

Locus	Mantel's z-statistic	Permutational s-score
LuTFL1	43.01*	24.27***
LuTFL2	15.35	20.60

Cultivated flax samples are mostly included in the haplotypes I and III of *LuTFL1* network. Additionally, there are three single samples that belong to the other pale flax haplotypes (VI, VIII, IX). Furthermore, there are additional six haplotype variants that were not found in pale flax (XI, XII, XIII, XIV, XV and XVI). These haplotypes are all associated with the haplotype I and III and hence will be referred to as cluster I and cluster III. No precise geographic data were available for the most of cultivated samples and hence centroids of the countries of origin were taken as a proxy of their latitude. The relationship between latitude and *LuTFL1* haplotype was modelled with logistic regression (Figure **4.1b**). Haplotypes were recorded as binary data; sample belong to either haplotype cluster I or III in *LuTFL1* network. The regression model fitted well with 99.9% confidence level (p-value of 0.00226). The increase of frequency of haplotype III with each grade of latitude was calculated to be 7.9% (between 2.9% and 13.5% with 95% confidence). This result confirms that *LuTFL1* haplotypes are correlated with latitudinal gradient in cultivated flax.

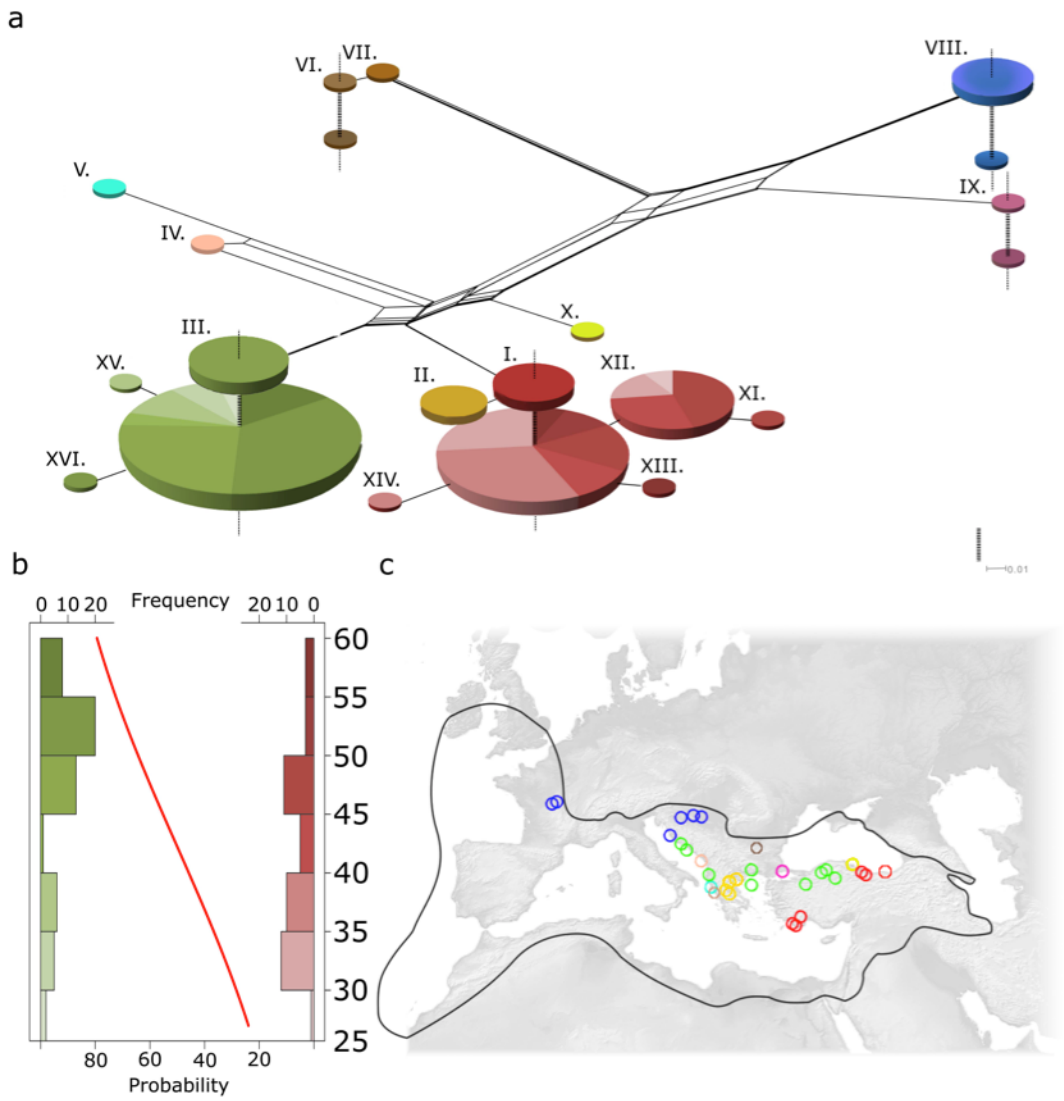


Figure 4.1: Relationship of cultivated and pale flax samples according to *LuTFL1* data and their geographic origins.
a - *SPLITSTREE* network of pale (top) and cultivated (bottom) flax, size of nodes is proportional to number of samples with the same haplotype, continuous branches denote molecular distance between, dotted lines link different species within same haplotype.
b – histogram showing latitudinal gradient in cultivated flax, to the left frequency of northern haplotype cluster (green), to the right frequency of southern haplotype cluster (red) with fitted logistic regression curve reflecting occurrence probability of northern haplotype in latitude gradient
c – map of Europe marked with wild distribution of pale flax (black line) and pale flax sampling locations (colours correspond to haplotypes in splits network).

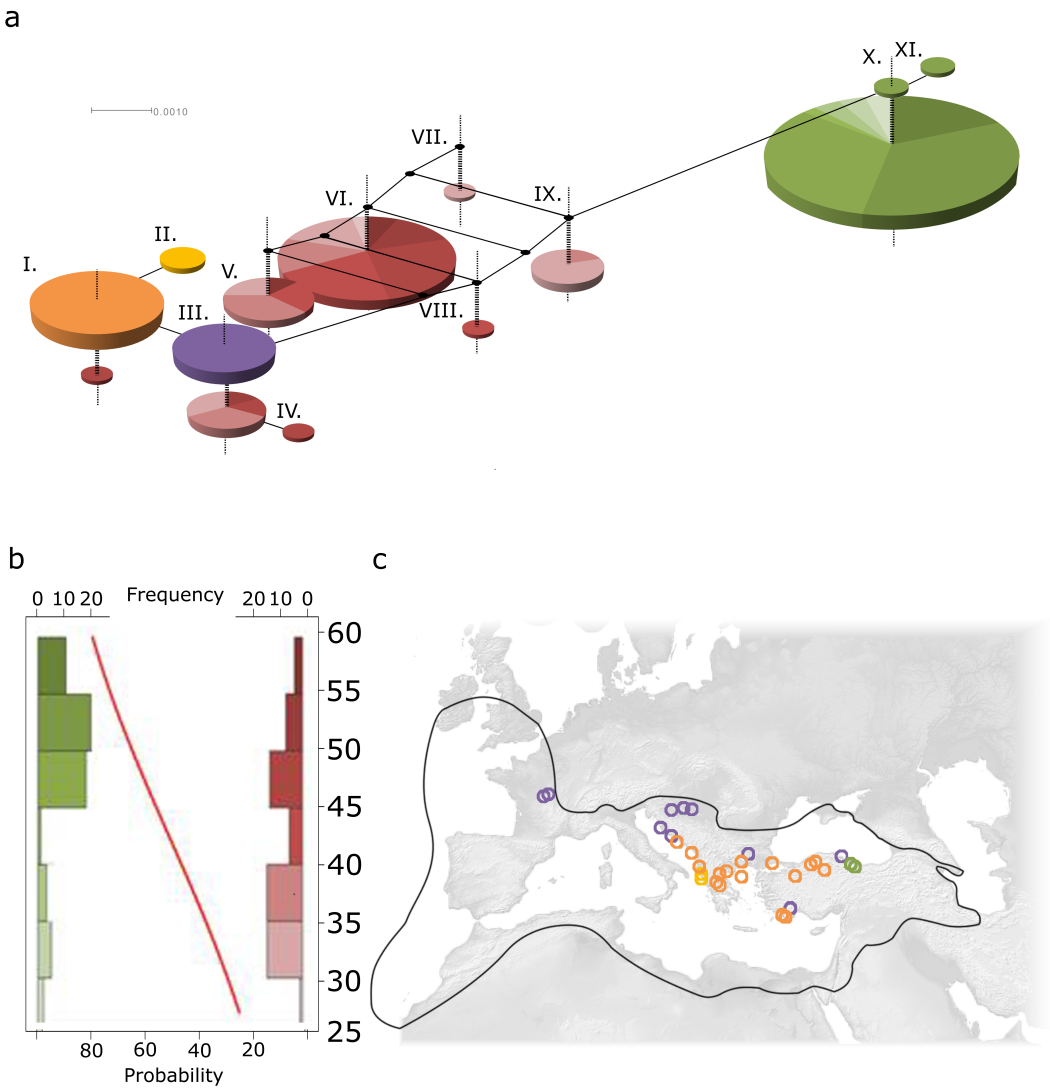


Figure 4.2: Relationship of cultivated and pale flax samples according to *LuTFL2* data and their geographic origins.
a - *SPLITSTREE* network of pale (top) and cultivated (bottom) flax, size of nodes is proportional to number of samples with the same haplotype, continuous branches denote molecular distance between, dotted lines link different species within same haplotype.
b – histogram showing latitudinal gradient in cultivated flax, to the left frequency of northern haplotype cluster (green), to the right frequency of southern haplotype cluster (red) with fitted logistic regression curve reflecting occurrence probability of northern haplotype in latitude gradient
c – map of Europe marked with wild distribution of pale flax (black line) and pale flax sampling locations (colours correspond to haplotypes in splits network).

By contrast, the *LuTFL2* based network contains only five haplotypes representing pale flax populations. Haplotype I with closely related haplotype II represent populations from Turkey and middle Balkans. All of the northern populations of pale flax are grouped together with two Turkish populations in haplotype III. Furthermore, there are two haplotype variants (X, XI) that contain single accession of pale flax each and are remote respective to former haplotypes, however, are only one mutation away from each other. Both pale flax samples originate from Northern Turkey. Despite a visible geographic pattern (Figure **4.2c**), *LuTFL2* haplotypes in pale flax are not significantly spatially autocorrelated based on incongruence of their geographic and genetic distances and permutational test (Table **4.3**).

Cultivated flax samples that are included in pale flax *LuTFL2* haplotypes I and III could reflect ancestry from southern and northern populations of wild flax respectively. One closely related haplotype variant stems out from haplotype III in cultivated flax (IV). A further five haplotypes occur exclusively in cultivated flaxes (V, VI, VII, VII and IX), with most accessions belonging to VI. This cluster of haplotypes consists of predominantly southern samples. The last and the largest haplotype is haplotype X representing mainly northern samples of cultivated flax and is clustered with two pale flax populations from north-eastern Turkey (Figure **4.2a** and **4.2c**). Again, when the southern haplotype cluster (I to IX) was compared to northern haplotype X, latitudinal correlation was established in cultivated flax. The regression model fitted well with 99.9% confidence level (p-value of 0.00029). The increase of frequency of haplotype X with each grade of latitude was calculated to be 10.3% (between 4.9% and 16.9% with 95% confidence). This result confirms that *LuTFL2* haplotypes also are correlated with latitudinal gradient in cultivated flax.

4.3.3 Linkage disequilibrium and signature of selection

Analysis of linkage disequilibrium between mutations in *LuTFL1* and *LuTFL2* resulted with different outcome depending on which of the two statistics (Lewontin's D' and squared allelic correlation r^2) was used. In both analyses mutations within *LuTFL1* and within *LuTFL2* separately were characterized by high scores and thus were probably inherited together. The difference between the two analyses was highlighted when linkage between *LuTFL1* and *LuTFL2* mutations was compared. Based on Lewontin's D' more than half mutations link both loci; LDE ranges from 0

to 1, with more than half scores above 0.5 (Figure **4.3**; Electronic Supplement **9**). By contrast, when different haplotype frequencies in the loci were taken into account, LDE scores were much lower; majority range from 0 to 0.24 (Figure **4.4**). Only in one case the score remained high ($r^2 = 0.57$); it characterizes linkage between mutation 686 in *LuTFL1* and 453 in *LuTFL2*.

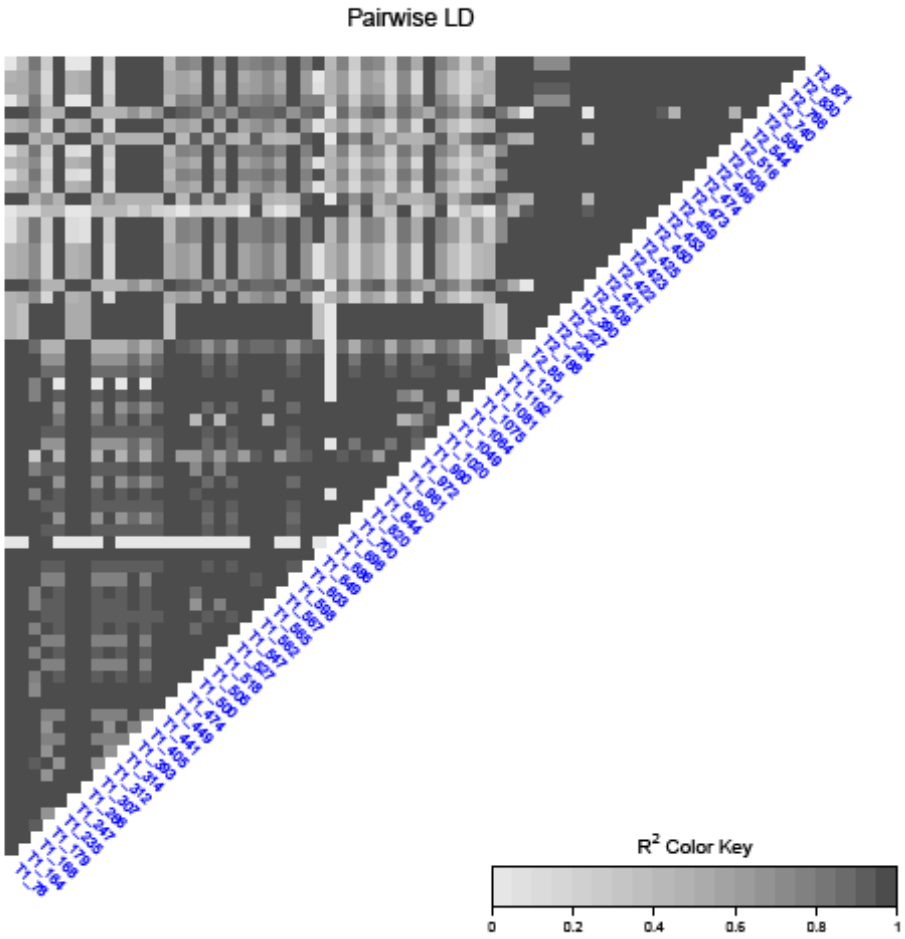
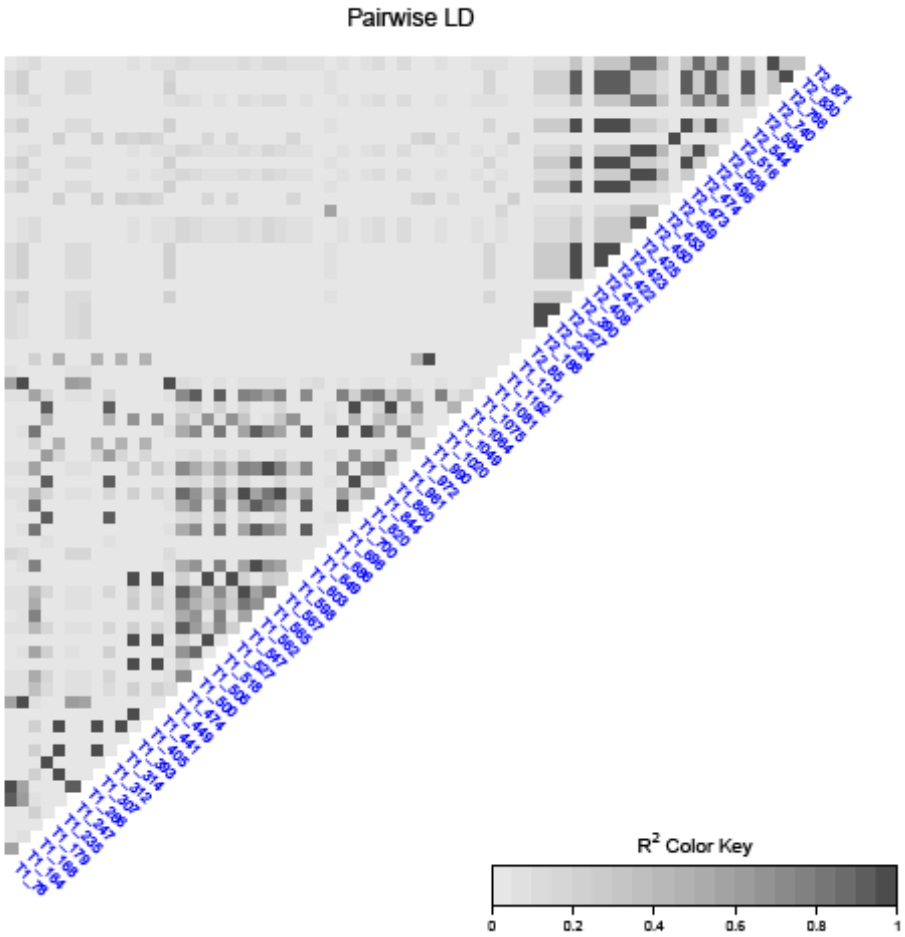


Figure 4.3: Heatmap reflecting LDE scores between mutations in *LuTFL1* and *LuTFL2* based on Lewontin's D'. Mutations in *LuTFL1* are named from T1_78 to T1_1211, while in *LuTFL2* – from T2_85 to T2_871. Legend: the grey scale gradient reflects the linkage between the two genes.



Fig

Figure 4.4: Heatmap reflecting LDE scores between mutations in *LuTFL1* and *LuTFL2* based on squared allelic correlation r^2. Mutations in *LuTFL1* are named from T1_78 to T1_1211, while in *LuTFL2* – from T2_85 to T2_871. Legend: the grey scale gradient reflects the linkage between the two genes.

LuTFL1 and *LuTFL2* were both analysed for neutrality of their evolution. Four different test methods were carried out for three datasets: pale, cultivated flax and combined data (Table **4.4**). The null hypothesis of neutral evolution could not be rejected for combined dataset in all the tests except for the R2 test in *LuTFL1*. Tests were also carried out for pale and cultivated flax separately. All four test methods resulted in rejection of the null hypothesis for the cultivated flax in *LuTFL1* locus. By contrast, tests for *LuTFL2* locus showed that all groups of flax evolved with no significant differences from neutrality. In pale flax none of the tests allowed for rejection of neutral evolution scenario.

Table 4.4: Results of neutrality tests for *LuTFL1* and *LuTFL2* loci.

Locus	Sample	Tajima's D test	Fu and Li's D_2 test	Fu and Li's F test	R2 test
LuTFL1	All	-0.7254	0.8105	0.2300	0.0375*
	Cultivated	-1.4701*	-4.6342**	-3.8800**	0.0401*
	Wild	-0.1403	0.0199	-0.0361	0.1099
LuTFL2	All	2.1806	0.1114	1.0638	0.1319
	Cultivated	1.5637	-0.0399	0.6442	0.1256
	Wild	-1.4443	1.2241	0.4099	0.0708

* rejection of null hypothesis with confidence of 0.05
** rejection of null hypothesis with confidence of 0.01

Table 4.5: Flax samples grouped according to their latitude of origin and frequencies of northern haplotypes in the *LuTFL1* locus.

Group	Latitudes	Number of samples	Frequency of north associated haplotype
1	50-55	20	0.82
2	45-50	13	0.65
3	40-45	9	0.22
4	35-40	20	0.38
5	30-35	34	0.35

Table 4.6: Peak of selection pressure required to observe 82% of advantageous haplotype after 12000 generations.

Initial haplotype frequency	Peak selection pressure						
	No migrants	0.1% migrants	1% migrants	2% migrants	3% migrants	4% migrants	5% migrants
0	-	0.07	0.062	0.05	0.036	0.03	0.028
0.3	0.03	0.03	0.03	0.028	0.028	0.028	0.028

The *LuTFL1* data were used to model northern movement of cultivated flax and its subsequent adaptation. Firstly, the samples used in this study were grouped according to the latitude of origin and their haplotype frequencies were calculated (Table **4.5**). The observed frequency that is associated with northern-most group equals 0.82. Computer simulations (for details see Section ***4.2.4***) were conducted and yielded spectrum of expected frequencies of advantageous *LuTFL1* haplotype in the north within 1000 replicates for each selection coefficient. Whenever the simulated

value matched the observed a point was scored. The highest scores were obtained when applied selection spanned between 0.01 and 0.1. Higher values of selection led to haplotype fixation, whilst lower – to removing it from gene pool through genetic drift.

Three different scenarios were tested in these simulations. In Scenario 1 advantageous haplotype existed within cultivated flax gene pool in the area of domestication and had initial frequency of 0.3 (Table **4.5**). In this scenario the highest score was obtained with selection pressure of 0.03 (Table **4.6**); this coefficient likely explains the distribution of observed data under scenario 1. By contrast, in Scenario 2 where the advantageous haplotype was introduced with wild gene flow, selection with highest scores ranged from 0.28 to 0.07 depending on number of migrants per generation. Selection has to be stronger with decreasing frequency of hybridization with wild species in order to reach frequency of 0.82 in the north. Scenario 3 assumed initial frequency of advantageous haplotype to be 0.3 within cultivated flax and allowed for admixture of advantageous haplotype from wild populations. In this case, peak of selection pressure was 0.3 (similar to that in scenario 1), however, it decreased slightly with increased migration to 0.28.

4.4 DISCUSSION

Despite the obvious morphological and phenological differences described by Diederichsen and Hammer (1995), cultivated and pale flax are very closely related suggesting ancestry of the domesticated species. In all of the domesticated crops we can estimate divergence time from its wild relative and region in which this divergence occurred using archaeological data. Primary domestication of cultivated flax took place in the Near East approximately 12,000 years ago; the earliest archaeological finds are recorded in Tell Ramad, Syria and dated back to 7,000 BC (Van Zeist & Bakker-Heeres 1975). This timeframe is essential in correct interpretation of haplotype networks and understanding the relationship between pale and cultivated flax based on *LuTFL1* and *LuTFL2* sequencing data.

4.4.1 The spread and hybridization of cultivated flax based on LuTFL1 locus

In the network based on *LuTFL1*, both species of flax share haplotypes I, III, VI, VIII and IX which suggests that gene flow occurred between them in the past or alternatively that we are observing an effect of lineage sorting (Figure **4.1a**). Haplotypes of *LuTFL1* are geographically structured in pale flax. For example, haplotype I occurs only in Turkey, while haplotype VIII in the northern Balkans. Cultivated flax was domesticated in the Near East and therefore could not have inherited haplotype VIII in any other way than through hybridization with northern pale flax. Furthermore, the gene flow between pale and cultivated flax is biologically possible as it was shown that these species are interfertile (Muravenko *et al.* 2003; Yermanos & Gill 1969).

The gene flow apparent from the *LuTFL1* haplotype network probably occurred from pale to cultivated flax. Pale flax samples are evenly distributed within this network and according to statistical tests their haplotypes do not show significant differences from the expected distribution under neutral evolution (Table **4.4**). By contrast, the cultivated flax network is highly biased. The most of the samples (103) are grouped into the two clusters of haplotypes (cluster I and cluster III), which are nested in nodes of their wild relatives, while another three specimens are sharing less common, distant haplotypes with pale flax. The difference in distribution of pale and cultivated flax haplotypes in this network suggests that *LuTFL1* haplotypes evolved within pale flax population and thus the direction of the gene flow was from the wild to the

cultivated species. The network in which such gene flow is assumed is more parsimonious then the network in which these haplotypes evolved in cultivated flax first and then were transferred onto pale flax. Morphological difference between both species is also counted as a modification in a sense that they also are taken into account when the most parsimonious solution is being chosen. Additionally, bearing in mind relatively recent divergence of cultivated flax it makes more sense to attribute haplotype diversity to long-existed wild species. Therefore, it is concluded that instances of gene flow pictured in *LuTFL1* network reflect original domestication event (haplotype I) and multiple instances of genetic influx from pale flax to cultivated flax as it spread northwards to Europe (haplotypes III, VI, VII and IX; Figure **4.1a**).

The *LuTFL1* haplotypes of pale flax show strong geographical structure. Haplotype I is exclusively associated with populations inhabiting south and north-eastern Turkey and is closely related to haplotype II occurring in the south-eastern Greece. Moving towards North, on both sides of Bosphorus strait haplotype III occurs predominantly. Northern Balkans is inhabited by haplotype VIII only. The minor haplotypes with only single representative except of IX fit this pattern very well. The distribution described above is associated with mixture of latitude and longitude, following diagonal gradient across Europe (Figure **4.1c**). Of all the populations occurring above the 40th degree of latitude only these from north-eastern Turkey retained haplotype I (associated with plants inhabiting Southern Europe). Interestingly, this population is more related to southern relatives then to adjacent north-western populations according to ISSR study as well (Uysal *et al.* 2010). Genetic isolation of both northern populations could be a result of geographic isolation - Koraoglu mountain range. Latitudinal pattern is very clear for haplotypes III and VIII. Coincidentally, the geographic structure associated with *LuTFL1* haplotypes reflects different stages of spread of Neolithic revolution to Europe.

Given that the *LuTFL1* locus was often transferred from pale to cultivated flax during secondary contacts of the both species and that its haplotypes are associated with latitude, we could use *LuTFL1* as molecular marker in which polymorphism reflects migration stages towards north. In cultivated flax haplotype I is enriched with samples from southern, whereas haplotype III is enriched in northern populations. Regression analyses imply that these haplotypes are significantly correlated with

latitude (Figure **4.1b**). Altogether, the data suggest that instances of past gene flows from wild to cultivated flax were associated with geography. Normally, gene flow between two mainly self-pollinating species would be difficult to observe in genetic data, perhaps geographic enrichment of these haplotypes in both pale and cultivated flax is a result of selection. This line of inference is continued in Section ***4.4.2***.

It is remarkable that haplotype XII is also enriched with northern accessions of flax. Interestingly, it belongs to predominantly southern cluster I. The mutation that discriminate haplotypes I and XII is a non-synonymous mutation in exon 4. It might be the case that this mutation caused another northern phenotype. It is concluded that perhaps haplotypes from cluster III and haplotype XII independently evolved adaptation to the northerly latitudes. Interestingly, the simple, non-synonymous mutation that occurred in haplotype XII is less frequent within northern populations than mutations associated with cluster III. This might reflect the advantage of long-standing, complex adaptations (cluster III) over recently evolved, simple mutations (haplotype XII).

4.4.2 The potential role of LuTFL1 in adaptation to northerly latitudes

The *LuTFL1* haplotype network within cultivated flax does not match the criterion of neutral evolution and hence could be driven by 1) recent population expansion 2) selection or 3) genetic hitchhiking. Because of the recent divergence time and rapid spread, a flax population expansion would seem the most plausible option, however, in this case we would expect to see pattern of non-neutrality in *LuTFL2* haplotypes as well. Both of these loci are very similar and have got conserved sequences on amino acid level. The great majority of polymorphic sites are in non-coding introns of these putative genes. Despite all that, *LuTFL2* evolution does not differ significantly from neutrality while *LuTFL1*'s does (Table **4.4**). For this reason rapid population expansion scenario was rejected. Alternatively, distribution of *LuTFL1* could be explained by an impact of selection or selective sweep. This locus is non-neutral in cultivated flax only, its wild relative's evolution has not deviated from neutrality. This incongruence could be explained with different distribution of the both species and the correlation between *LuTFL1* haplotypes and latitude. If the nature of selection on *LuTFL1* was latitudinal then it would impose larger pressure to populations with geographic distribution stretching further north. By contrast to

unconstrained distribution of cultivated flax in Europe, its wild relative occurrence is restricted mostly to Southern Europe with the exception of the southern shore of the UK and Ireland (Vul'f & Elladi 1940), hence the selection pressure would be much weaker in the latter. Furthermore, the spread of cultivated flax was rapid on evolutionary time scale and hence the applied selection might have been strong. Therefore, it is concluded that direct selection or selective sweep could have been applied to both species at different strengths and thus allowed for detection in cultivated flax only.

Characterizing the selection type and estimating the time when it was applied to cultivated flax is very difficult. Abnormality in evolution of *LuTFL1* is not an effect of recent breeding practices as more than 30% of samples used in this study are historic (predate World War II; see Tables **2.1**, **2.2**, Supplement **1**). Hence, the only remaining explanation is that selection or genetic hitchhiking effect associated with *LuTFL1* haplotypes occurred historically. In case of *LuTFL1* locus it is difficult to apply methods of distinguishing different types of selection (reviewed in Nielsen 2005), as they require data for two different species that do not hybridize. The only way to have an insight into this issue is by comparing numbers of heterozygotes between the two analysed loci. Lower proportion of heterozygotes (around 8%) in *LuTFL1* locus when compared to *LuTFL2* (around 12%) might be suggesting disruptive selection in the former. This is congruent with the scenario in which *LuTFL1* I was selected for in the south while *LuTFL1* III in the North. Furthermore, geographic structure and limited migration between southern and northern cultivated flax might be indicative of local adaptations (Charlesworth *et al.* 1997) and hence forces that shaped genetic variation in *LuTFL1* locus probably reflect the stage of spread and adaptation of flax rather than domestication or yield improvement stages.

Genotyping *LuTFL1* gene in archaeological samples would allow to test directly if mutations in haplotype III were selected for in Central Europe and had an impact on decreasing flax seed size in the early Neolithic. Unfortunately, utility of DNA extracts from archaeological samples of flax is very low. Archaeological material of flax seeds from Neolithic Alpine Dwellings (Herbig & Maier 2011) was tested. It was impossible to amplify 40 bp long DNA fragments in a PCR due to heavy inhibition. When aDNA extract was added to modern genomic DNA (at concentration of 20ng/µl) visible inhibition of PCR was observed. This is probably

due to small size of a single flax seed, which decreases DNA preservation. Furthermore, archaeological samples from Alpine dwellings of Neolithic Central Europe were preserved by waterlogging. Presence of water increases DNA degradation through depurination.

The *TFL* gene family is an important factor in regulation of flowering time and growth determinacy, and thus could play a role in adaptations to shorter vegetative periods associated with the north. Archaeological data suggest that in Central Europe summer crops were grown by contrast to winter varieties predominantly cultivated in the Balkans (Kreuz *et al.* 2005). According to Purugganan and Fuller (2009) winter crops needed to adapt through changed vernalization response or photoperiod neutrality. In barley close homolog to *TFL – HvCEN* was shown to play an important role in adaptation to European latitudes through change from winter to spring growth habit. All major haplotypes were shared between wild and cultivated plants, however, in this case no geographic association was established from wild barley (Comadran *et al.* 2012). In flax a similar pattern is observed, where *LuTFL1* haplotypes (*CEN* homolog) are associated with latitudinal gradient, however here, evidence is presented that haplotypes were passed from geographically distinct pale flax populations. It might be the case that in cultivated flax *LuTFL1* deviated from neutral evolution due to latitudinal selection during the spread in Europe.

Computer simulations support the notion that *LuTFL1* was under selection or selective sweep; it is the only explanation for observed haplotype frequencies. In all scenarios application of selection is essential for haplotype III to reach observed frequency, without it haplotype is lost due to genetic drift, even with influx from wild populations. Interestingly, the selection pressure that yields highest scores is not significantly impacted by rates of gene flow between wild and cultivated flax in scenario 2 and 3. In this study, simulations with 0.1 to 5% of immigrants have been tested; this numbers are based on natural crosspollination limits of cultivated flax (Gurbuz 1999) and migration data from self-pollinating barley (Hubner et al. 2012). Regardless of whether there was 1 or 50 migrants from pale flax to population of 1000 cultivated plants, the highest score selection pressure remained lower than 0.1. That shows that when selection is applied even low rates of gene flow will enable acquisition of an advantageous haplotype from another population rendering the *LuTFL1* III transfer scenario possible. Furthermore, the simulations suggest that

different scenarios of origin of adaptation require different selection pressure to get a maximum score. There is, however, no data on strength of selection in similar cases so we cannot conclude, which scenario is more likely based on simulations output.

4.4.3 The spread of cultivated flax based on LuTFL2 locus

In the network based on *LuTFL2* locus lower genetic diversity within pale flax can be observed than that for *LuTFL1* (Figures **4.1a** and **4.2a**). There are only five haplotypes representing pale flax. Haplotypes I, II and III, each separated with one mutation are forming the core cluster. On the other hand, there are haplotypes X and XI, each with single pale flax accession – these form the satellite cluster. There are four samples of cultivated flax that share haplotypes with the core cluster and 51 samples that share haplotype X with satellite cluster. These two clusters are separated by haplotypes that are exclusive to cultivated flax (IV, V, VI, VII, VIII, IX). Samples of cultivars represent much higher molecular diversity than these of pale flax, which is exactly the opposite to the situation observed in haplotype network for *LuTFL1*. In consequence, *LuTFL2* network is more parsimonious if we assume that gene flow in core cluster represents domestication but the current of gene flow in haplotype X was from cultivated to pale flax. This solution however, does not make much sense when geographic origin of cultivated and pale flax in satellite cluster is taken into account.

There is an unclear pattern in the geographic distribution of haplotypes within pale flax based on *LuTFL2* data (Figure **4.2c**). All the northern samples and additionally two samples from Turkey belong to haplotype III. Samples from both sides of Bosphorus strait are grouped mainly in haplotype I while two western-most Greek samples belong to haplotype II. Finally, the two eastern-most Turkish samples form haplotypes X and XI. Based on statistical tests however, haplotypes are not spatially autocorrelated (Table **4.3**). The most of cultivated flax samples are grouped in haplotype X. Paradoxically however, they are heavily enriched with samples from the northerly latitudes. It is difficult to imagine gene flow between eastern Turkish pale flax and Northern European cultivated flax. Therefore, the situation in which the two species share haplotype X could be explained with an effect of lineage sorting. All of the pale flax haplotypes are present in the area of domestication and therefore it might be the case that Northern cultivated flax inherited haplotype X during domestication.

Pale flax seems not to be spatially autocorrelated based on *LuTFL2* data, however, there is a visible enrichment of certain haplotypes with changing latitude in cultivated samples (Figure **4.2b**). Based on correlation study northern populations of cultivated flax are much more likely to carry haplotype X, while populations in the south have high possibility of carrying other haplotypes. This is congruent with the interpretation in which core cluster of pale flax (haplotypes I, II and III) contains ancestors of cultivated flax and that the shared haplotypes within this cluster reflect domestication event. Remaining questions regard the cause of the latitudinal gradient in *LuTFL2* haplotypes and if it played any role in adapting cultivated flax to the north.

The geographic distribution of *LuTFL2* haplotypes cannot be explained by the presence of selection. It was stated in Section ***4.4.2*** that the observed molecular diversity in *LuTFL2* haplotypes could be a result of neutral evolution. The hypothesis of neutrality could not be rejected based on any of the tests carried out in this project. It seems unlikely that latitudinal correlation of *LuTFL2* haplotypes could be an artefact. Furthermore, in pale flax isolation by distance was ruled out, since geographic distances were not congruent with genetic distances between samples based on Mantel's test (Table **4.3**). The remaining solution is that *LuTFL2* locus is linked and inherited together with another locus that is under selection, however, this link is too weak to leave a signature of selective sweep in the *LuTFL2* data. For this reason, linkage disequilibrium test was carried out to see if mutations in *LuTFL2* are inherited congruently with mutations in *LuTFL1*. The high LDE scores based on Lewontin's D' are probably false positive; they could be a result of not taking haplotype frequencies in both loci into account (Figure **4.3**). Nevertheless, there are still mutations that are in LDE between the two *LuTFL* loci based on squared allelic correlation (Figure **4.4**). The most notable case is LDE between mutation 686 in *LuTFL1* and 453 in *LuTFL2* with the score of 0.57. Mutation 686 is a 20 bp long indel right after exon 2 which characterise haplotype VIII (northern-most pale flax and one cultivated flax sample), while mutation 453 is a substitution discriminating haplotype I (southern pale flax) from haplotype III (northern pale flax). This strong link between mutations in *LuTFL* loci however, seem to be only present in pale flax. In cultivated flax, LDE between the mutations in the two loci is weak, congruently with lacking signature of selective sweep in *LuTFL2*. The conclusion drawn from

these analyses is that LDE between *LuTFL2* and *LuTFL1* in cultivated flax is too weak to detect genetic hitchhiking, however, might have been strong enough to influence geographic structure of haplotypes in *LuTFL2*. Alternatively, *LuTFL2* might be weakly linked to another locus, which haplotypes present latitudinal correlation.

4.4.4 Summary

The above findings provide an explanation of how cultivated flax has changed during its spread from the Near East to Central Europe. Since geographic distribution of wild flax stretches far outside the Near East, we can assume that its gene pool accommodates spectrum of adaptations to different geographic regions of Europe and Asia. Both molecular and phenotypic data suggest presence of strong local adaptations in Turkish populations of pale flax (Uysal *et al.* 2010; Uysal *et al.* 2012). These adaptations could have been transferred onto cultivated flax through hybridization. Even narrow species bridge could lead to increasing frequency of a new allele due to selection pressure. In this scenario we assume that pale flax served as genetic reservoir and could have transferred adaptations to its domesticated relative possibly resulting with major changes in flax phenology and/or morphology or perhaps even with new varieties. Pale flax haplotypes of *LuTFL1* might have played an important role in adaptation of cultivated flax to northerly latitudes through change in flowering regime and/or growth determinacy. At the same time, *LuTFL1* is a homolog of *TFL1* gene in *A. thaliana* and as such potentially have an impact on plant height and architecture, which in turn might have led to emergence of fibre varieties in Central Europe. The potential of *LuTFL1* in changing flax phenology and architecture is explored in **Chapter 5**.

CHAPTER 5: EXPRESSION ANALYSES OF *LuTFL1*, *LuTFL2*, *LuFT* AND THEIR IMPACT ON FLAX PHENOLOGY AND STATURE

5.1 INTRODUCTION

Cultivated flax arrived to Central Europe around 5000 years BCE and has undergone major morphological changes. There is archaeological evidence for the decreasing importance of linseed and for specialization in fibre production (Herbig & Maier 2011). The two specialized varieties differ in size of seeds, number of floral branches and plant height. The change in flax morphology might have been linked to the adaptation to northerly latitude through change in flowering regime. Numerous aspects of crop plant morphology are affected by different flowering strategies. Among the most important are: mode of growth (sympodial vs. monopodial), plant height, number of flowers and quality of fruits produced (Elitzur *et al*. 2009; Krieger *et al*. 2010; Shalit *et al*. 2009). These associations create a link between latitude and plant morphology. For example, poplar phenology genes show elevated phenotypic differentiation along a latitudinal gradient for traits such as timing of the bud set and height growth (Keller *et al.* 2011b). All of these are very important agronomic traits and hence are of interest to both evolutionary biologists and plant breeders. Therefore, the association between flowering time and agronomic traits could be the key to understanding the process of the emergence of flax fibre varieties.

The aim of this part of project is to study expression of selected PEBP homologs in pale flax and investigate their potential impact on flax architecture and flowering time. In order to carry out phenotyping and expression analyses a time course experiment was designed. Additionally, an effort was made to model *FT*/*TFL* expression, its influence on flowering time and impact on height, number of floral branches and architecture of cultivated flax. This model was then confronted with the observed data.

5.2 MATERIALS AND METHODS

5.2.1 Phenotypic analyses in pale and cultivated flax

A time course experiment was set up for phenotyping and expression analyses in selected accessions of pale flax. Plants of six populations (W042, W043, W067, W069, W077 and W094), representative of three geographical regions (Turkey, Greece and Croatia) were chosen to form the core experimental plants and sown in 100 replicates each. These plants were used for the gene expression experiment. Additionally 26 remaining populations of pale flax (full list is presented in Table **2.4**) were sown in five replicates for phenotyping experiment. All plants were sown in trays containing 40 pots with general compost. In order to equalize germination time for all the seeds, they were watered and kept in 4°C in darkness for three days covered with film. For the subsequent ten days, they were grown in custom made growth cabinet at room temperature with light set up for 16 hours. These cabinets were moved to a cold room at 4°C a vernalization period of 40 days. During this period six individuals in three replicates from within the core experimental plants were moved to 24°C after 10, 15, 20, 25, 30 and 35 days to assess minimum required vernalization time. All the other plants were moved after 40 days to MLR-351 Versatile plant growth (SANYO Electric Co., Ltd.) cabinets set up to emit light for 16h in temperature of 24°C during a day and 16°C during a night. Plants were grown in these conditions for another 40 days and then moved to the larger pots and into the glasshouse (daylength between 16 h 42 min and 15 h 55 min). Flowering time was recorded when floral primordia appeared for the first time. After maturation, plant height was measured along with weight of 1000 seeds for each pale flax population. The data for Turkish populations were obtained from authors of phenotypic study in Turkish pale flax (Uysal *et al.* 2010). Additionally, data generated by our collaborators from PGR Canada (Diederichsen 2009; Diederichsen & Raney 2006; Diederichsen & Ulrich 2009) for 1146 accessions of cultivated flax were used in this study (Electronic Supplement **10**). Correlation between mass of 1000 seeds, height, fibre content, branching pattern and flowering time was measured in *R* using Pearson's and Kendall's ranks.

5.2.2 Analysis of expression of selected PEBP genes in pale flax

During the time course experiment, at increasing intervals (0, 15, 17, 19 and 19 days) samples of three plants were taken from each of the core pale flax accessions. Sampled material included the main meristem, 1 cm of stem from the meristem and adjacent leaves. Tissue was rapidly frozen in liquid nitrogen and stored at -80°C. Frozen tissue was then ground in a pre-chilled mortar on dry ice. The total RNA extraction was carried out from 20 mg of ground tissue using mirVana™ miRNA isolation kit (Invitrogen) following the standard protocol. In order to check if RNA was intact, the sample was subject to electrophoresis in a 1% agarose TAE gel with ethidium bromide and compared against a 100 bp DNA ladder. RNA concentration was measured in a NanoDrop spectrophotometer. DNA contaminants were digested in reaction with 2 units of DNase I (Invitrogen) and 2 μg of total RNA in DEPC-treated water.

Table 5.1: Primers used in specific amplification of cDNA for semi-qPCR analysis and their melting temperatures.

Gene	Primers	Melting temperature
LuTFL1	AGTGATTGGGAGAGTAATAGGAG TTAGGCATGTGGGAAACAAGA	56°C
LuTFL2	AGTGATAGGGAGAGTGATTGGAG TTAGGCTTCTGAGTAACAGCAGA	59.5°C
LuFT	TCAGTTGGGATACACCGTTTC AGACGGAAGCAACAGGCG	56°C
GAPDH	AGGTTCTTCCCGCTCTCAAT CCTCCTTGATAGCAGCCTTG	58°C

Further to that, total cDNA was synthesized from 1μg of DNase-treated RNA with use of SuperScript® II reverse transcriptase and unspecific Oligo dT primer following the manufacturer instructions. The concentration of each cDNA sample was adjusted to 100 ng/μl. Specific primers were designed in exonic regions of *LuTFL1*, *LuTFL2* and *LuFT* to cover approximately 150 bp long fragment (Table **5.1**). Based on the recommendation from the literature *GAPDH* housekeeping gene was chosen as an expression control (Huis *et al.* 2010). Semi-quantitative PCR was carried out for each sample with 30 cycles. Each cycle consisted of the following

steps: 30 s in 92°C for denaturation, 30 s in 56-59.5°C for annealing (Table **4.1**) and 30 s in 72°C for elongation stage. PCR products were visualized in UV on 2% agarose gel after electrophoresis with size marker. Two splice variants of *LuTFL1* were amplified using primers described in Table **4.1** with the same cycling conditions. This experiment was carried out for two accessions (W42 and W42) in 7 time intervals. High-resolution electrophoresis for splice variants of *LuTFL1* was carried out in 4% agarose gels. Purified product was then subject to sequencing.

5.2.3 Floral architecture development model

A script in the *R* programming language called *PGROWTH* has been written by the author of this thesis to simulate the impact of different expression levels of *TFL1* and *FT* homologs on flax architecture during its development (source code attached as Electronic Supplement **11**). It is a mathematical implementation of observations made by McGarry and Ayres (2012b) on the impact of *FT*/*TFL* expression on tomato plant architecture. The type of growth and inflorescence architecture characteristic of flax were taken into account when modelling the development. The model assumes an indeterminate and uniform growth of stems and adjacent leaves over time. In the scenario tested, leaves were produced every five days. Plants were grown from beginning of March for 150 days. In the background, daylength is calculated using an empirical formula (Kirk 1994) based on current date in radians (β) and input latitude in radians (α):

$$tan(\alpha) * tan\left(-\frac{\pi(22.9133*cos(\beta)+4.02543*sin(\beta)-0.3872*cos(2*\beta)+0.052*sin(2*\beta))}{180}\right) * \frac{180}{\pi}.$$

The most important input parameters in this model are the daylength threshold that up-regulates *FT* expression and relative proportion of *FT* to *TFL* proteins that induce inflorescence development. When a plant reaches the generative stage the rates of stem growth remains the same, however instead of producing leaves every fifth day, the main stem produces leafy-bracts and floral branches in turns. On the axillary stems leafy-bracts and pedicels are generated. Overall, this models a racemose corymb that most closely reflects flax inflorescence. The script includes graphical functions to visualize output architecture. Furthermore, a wrapper script has been

written in *R* by the author of this thesis to explore range of parameters in multiple *PGROWTH* simulations and summarize important traits such as flowering time and plant height. In this project the following parameters were explored: latitude from N30° to N60°, *FT* expression threshold set to 14 hours and *FT*/*TFL* ratio from 0 to 100 which should cover all possible architectures.

5.3 RESULTS

5.3.1 Phenotypic analyses of pale and cultivated flax

All the seeds collected during Balkan expedition (Table **2.3**) were regenerated successfully except for W066, W082 and W096 and therefore no data for these accessions is presented. In pale flax, vernalization requirement was measured. No plants flowered after 10 and 15 days of cold temperatures. After 20 days only one in six accessions flowered, while four flowered after 25 days. All plants flowered after 30, 35 and 40 days of vernalization. Mass per 1000 seeds was measured in all the pale flax plants together with flowering time. Plant height was successfully measured in only six core accessions (W042, W043, W067, W069, W077 and W094). Phenotypic data was combined with the data for Turkish pale flax populations (Uysal *et al.* 2012) and presented in Figure **5.1**. There is no significant correlation of any traits with latitude and the associations between morphological and phenological characters are not significant except for plant height, flowering time and seed mass (Figure **5.1**).

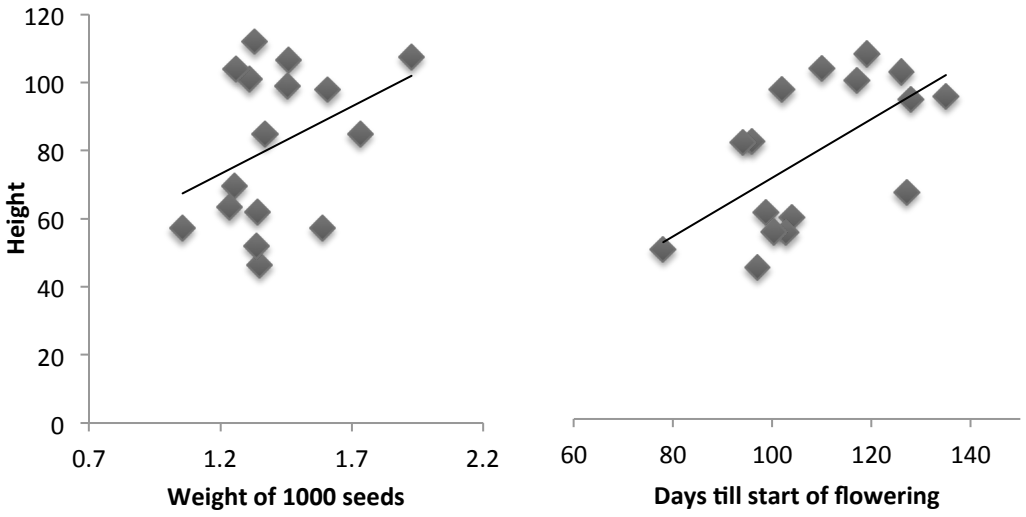


Figure 5.1: Correlation between plant height, seed weight and flowering time in pale flax. Plant height was measured for the tallest stem. Start of flowering was recorded on appearance of first floral primordium.

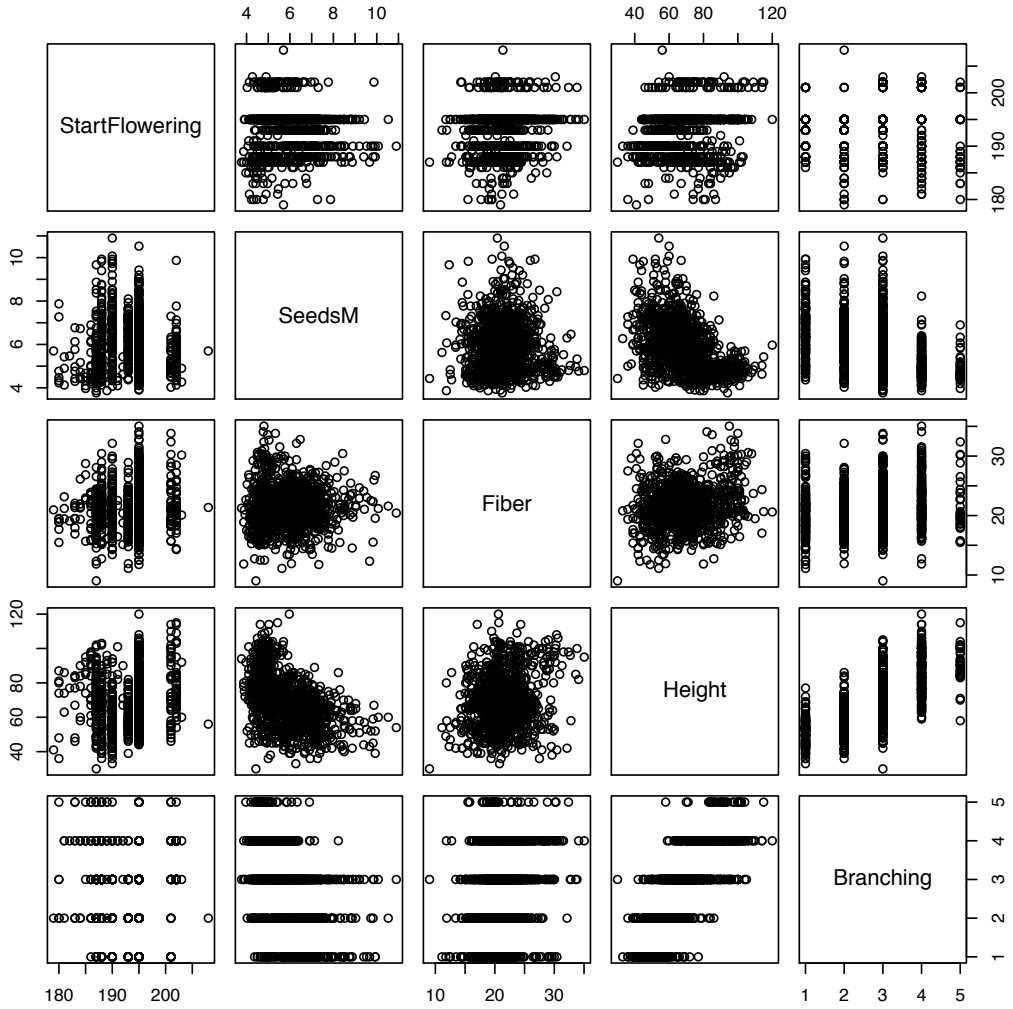


Figure 5.2: Scatter plots presenting pairwise association between flowering time and morphological traits in cultivated flax. Between investigated morphological traits were: mass of 1000 seeds, fibre content, overall plant height and branching pattern displayed as a proportion of branched stem length to not branched.

Table 5.2: Correlation coefficients between the start of flowering time and morphological traits in cultivated flax.

Rank correlation	Mass of 1000 seeds	Fibre content	Plant height	Branching pattern
Pearson's	-0.055	0.116***	0.166***	0.107***
Kendall's	-0.054*	0.075***	0.143***	0.161***

* result with significance level <0.05
** result with significance level <0.01
*** result with significance level <0.001

The association study in cultivated flax based on phenotypic and phenological traits in 1146 accessions revealed significant, positive correlation of flowering time with plant height and branching pattern, on the other hand – less significant, negative correlation with seed mass (Table **5.2**, Figure **5.2**). Additionally, plant height is negatively correlated with seed mass while positive correlation characterizes its relation with branching pattern.

5.3.2 Analysis of expression of selected PEBP genes in pale flax

The second part of time-course experiment in pale flax was carried out to assess expression patterns within genes *GAPDH*, *LuTFL1*, *LuTFL2* and *LuFT*. The expression strength of these genes is shown in Figures **5.3**, **5.4** and **5.5** for six core accessions. The housekeeping gene *GAPDH* was expressed in all the pale flax plants uniformly, except for the sample taken on day 55 in W042 plants. This sample was rejected from further analyses. *LuTFL1* was expressed in all the plants, with small differences in the strength of expression over time, this can be observed especially in W042 and W043. High-resolution, gel electrophoresis allowed distinguishing two variants of *LuTFL1* (Figure **5.6**). After sequencing, it was observed that the shorter variant is spliced without intron 3. Both splice variants were present in W42 and W43 accession invariably over all time intervals. By contrast to *LuTFL1*, *LuTFL2* was not expressed in the studied tissues at any time point analysed here. Finally, the greatest variation was observed in expression of *LuFT* gene. In all the plants it was expressed in 110th day, while in W042, W067 and W094 it was expressed already at day 91.

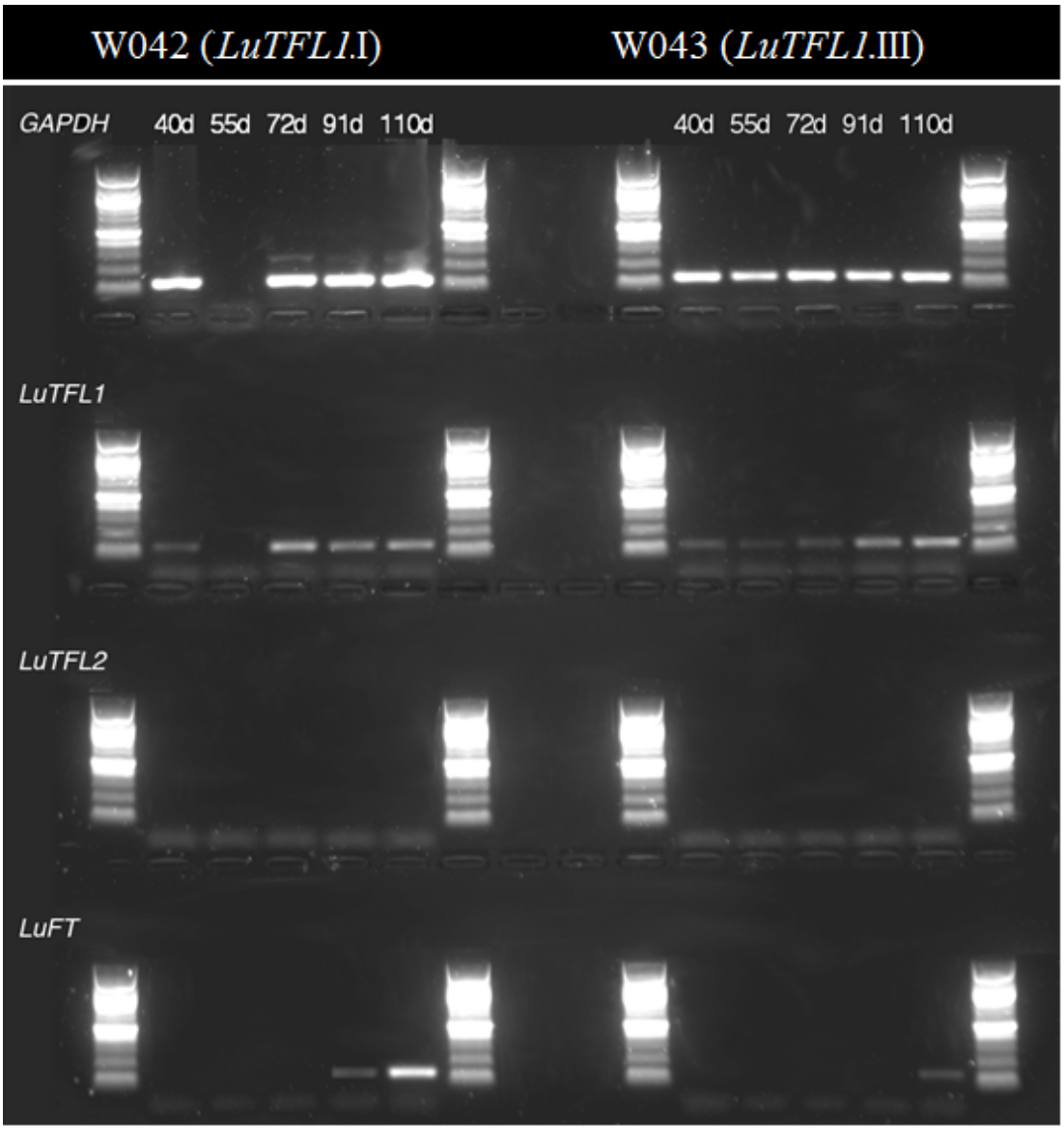


Figure 5.3: Electrophoresis gel showing the expression of PEBP genes in accessions W042 and W043. In brackets is the information about *LuTFL1* haplotype. Semi-quantitative PCR was carried out for *GAPDH*, *LuTFL1*, *LuTFL2* and *LuFT* loci based on cDNA template. Expression analyses were conducted in five time intervals.

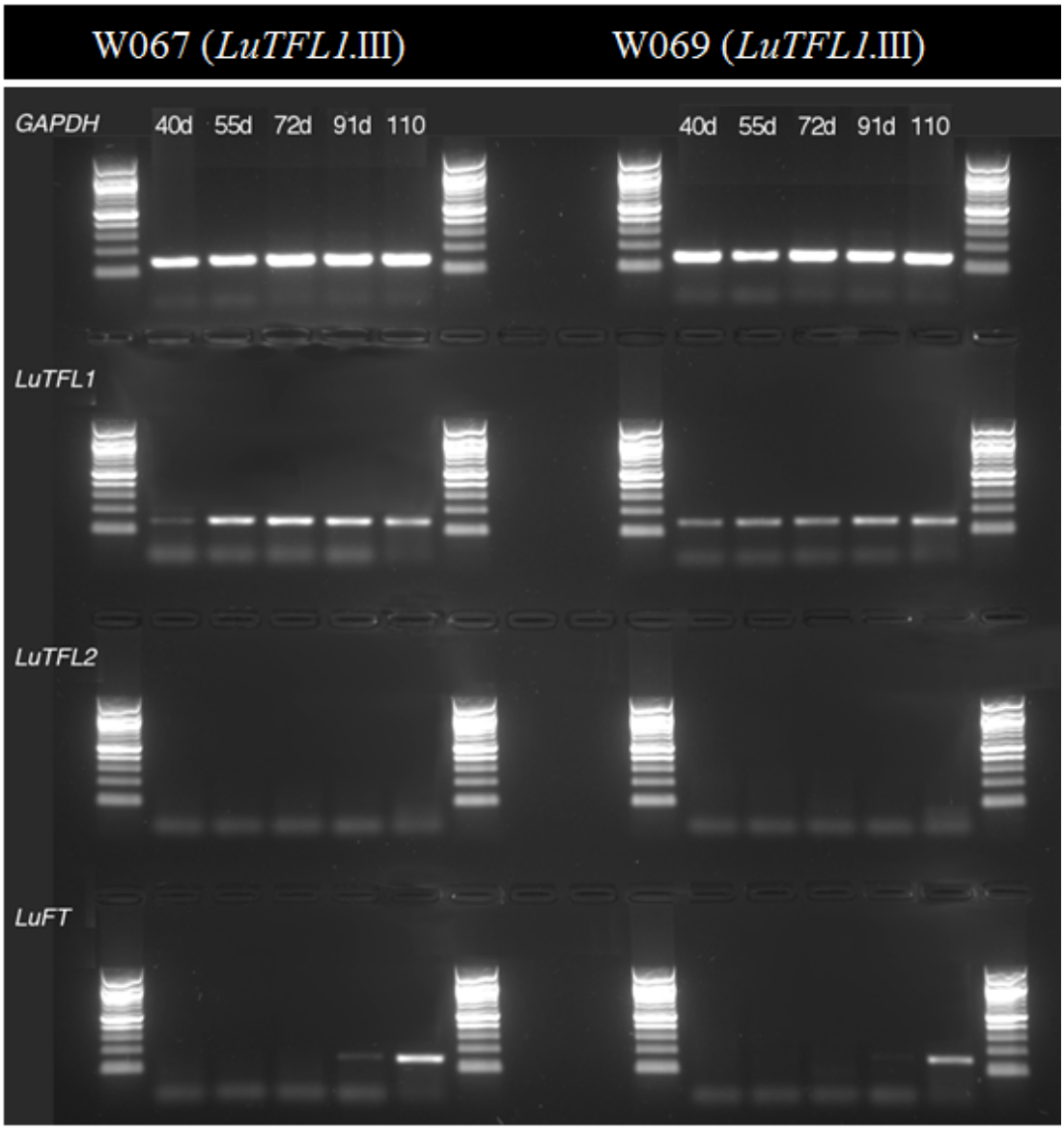


Figure 5.4: Electrophoresis gel showing the expression of PEBP genes in accessions W067 and W069. In brackets is the information about *LuTFL1* haplotype. Semi-quantitative PCR was carried out for *GAPDH*, *LuTFL1*, *LuTFL2* and *LuFT* loci based on cDNA template. Expression analyses were conducted in five time intervals.

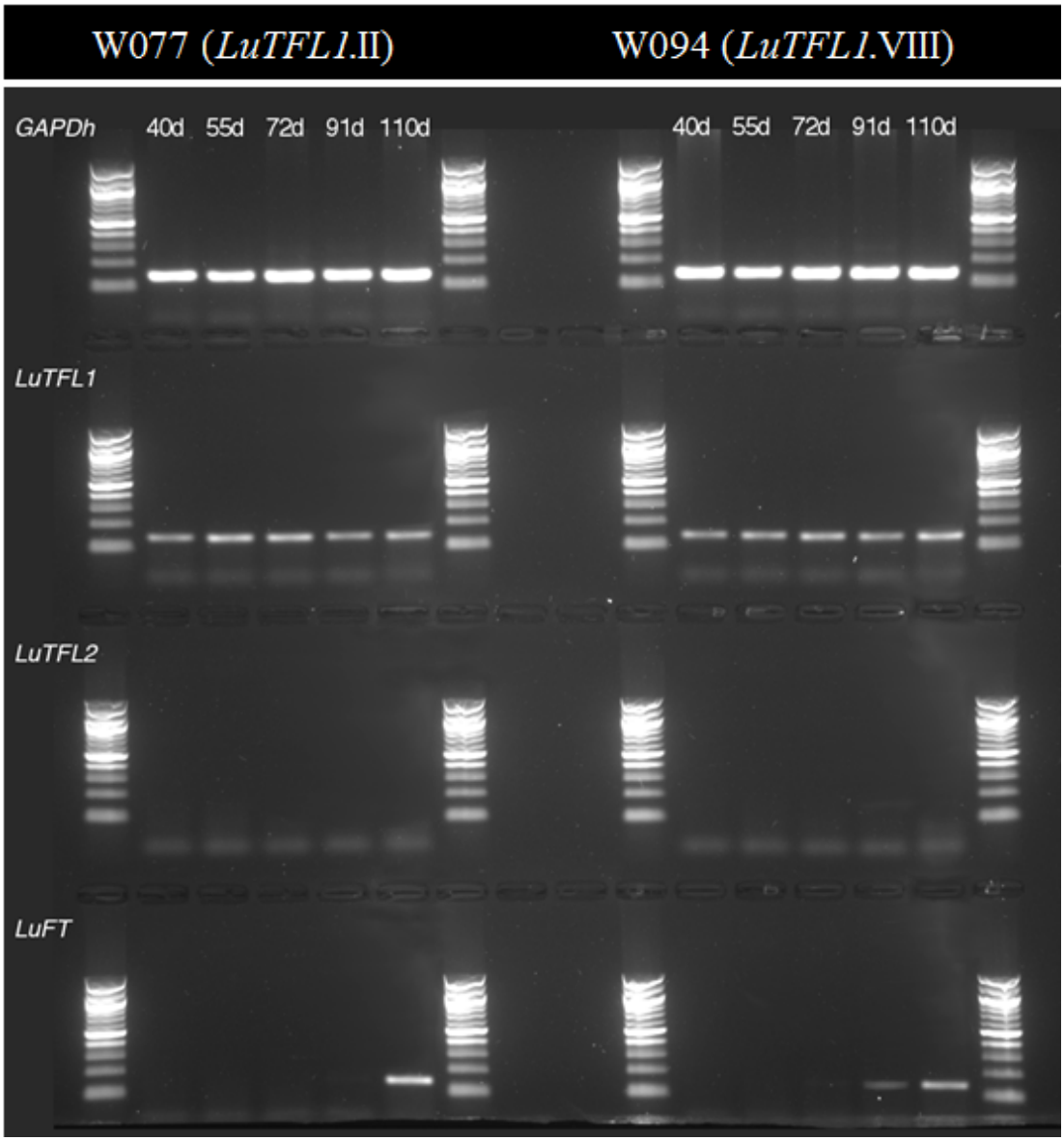


Figure 5.5: Electrophoresis gel showing the expression of PEBP genes in accessions W077 and W094. In brackets is the information about *LuTFL1* haplotype. Semi-quantitative PCR was carried out for *GAPDH*, *LuTFL1*, *LuTFL2* and *LuFT* loci based on cDNA template. Expression analyses were conducted in five time intervals.

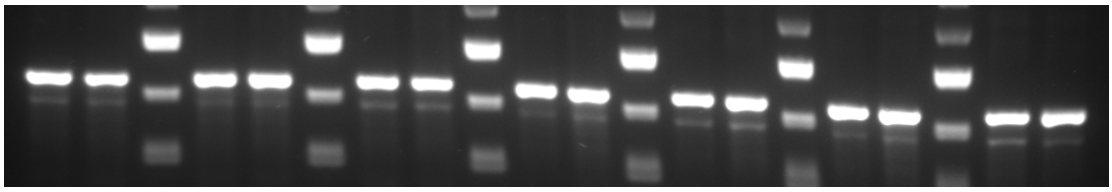


Figure 5.6: Electrophoresis gel showing the expression of two splice variants of the *LuTFL1* gene. Semi-quantitative PCR was carried out for *LuTFL1* based on cDNA template. High-density (4%) agarose gel was used.

5.3.3 Floral architecture development model

The *PGROWTH* model simulates flax growth and development. The key event determining plant architecture is flowering. Change in the flowering time has a huge impact on plant stature and branching pattern. The user-defined variable within this model is the ratio of the two major integrators of flowering (*FT* and *TFL*) that are necessary to trigger floral development. *TFL* is constantly expressed while *FT* is expressed only when the daylength threshold is met. Further vegetative growth boosts *FT* expression following the model suggested by McGarry and Ayre (2012b). When *FT*/*TFL* ratio reaches the set value it promotes flowering. These rules were applied in the development model for cultivated flax. Preliminary simulations were carried out with the *FT* expression threshold set to 14 hours and latitude set to N 40°. The range of *FT*/*TFL* ratio was explored (from 0.1 to 10). Different input settings yielded different inflorescence architecture; see Figure **5.7** for comparison of selected modelled architectures and sketches of different flax varieties made by Kulpa and Danert (1962). Most of the resultant architectures were similar to these observed in nature.

In the second part of the modelling experiment an impact of *FT*/*TFL* ratios on number of produced flowers, plant height until branching and total plant height was investigated in the latitudes spanning from N 35° to N 60°. These simulations of flax floral development showed the impact of both the latitudes and *FT*/*TFL* ratios on flowering time and on the proportion of height of main stem with floral branches to total height of a stem. There are two key results in this section. With increasing latitudes 1) lower *FT*/*TFL* ratios enable plants to flower earlier in a year, which in turn causes the short phenotype and 2) higher *FT*/*TFL* ratios enable plants to flower later in the calendar, which causes tall phenotype (Figure **5.8**). The conclusion drawn from these simulations is that more extreme architectures are allowed in the north due to higher amplitude of daylength.

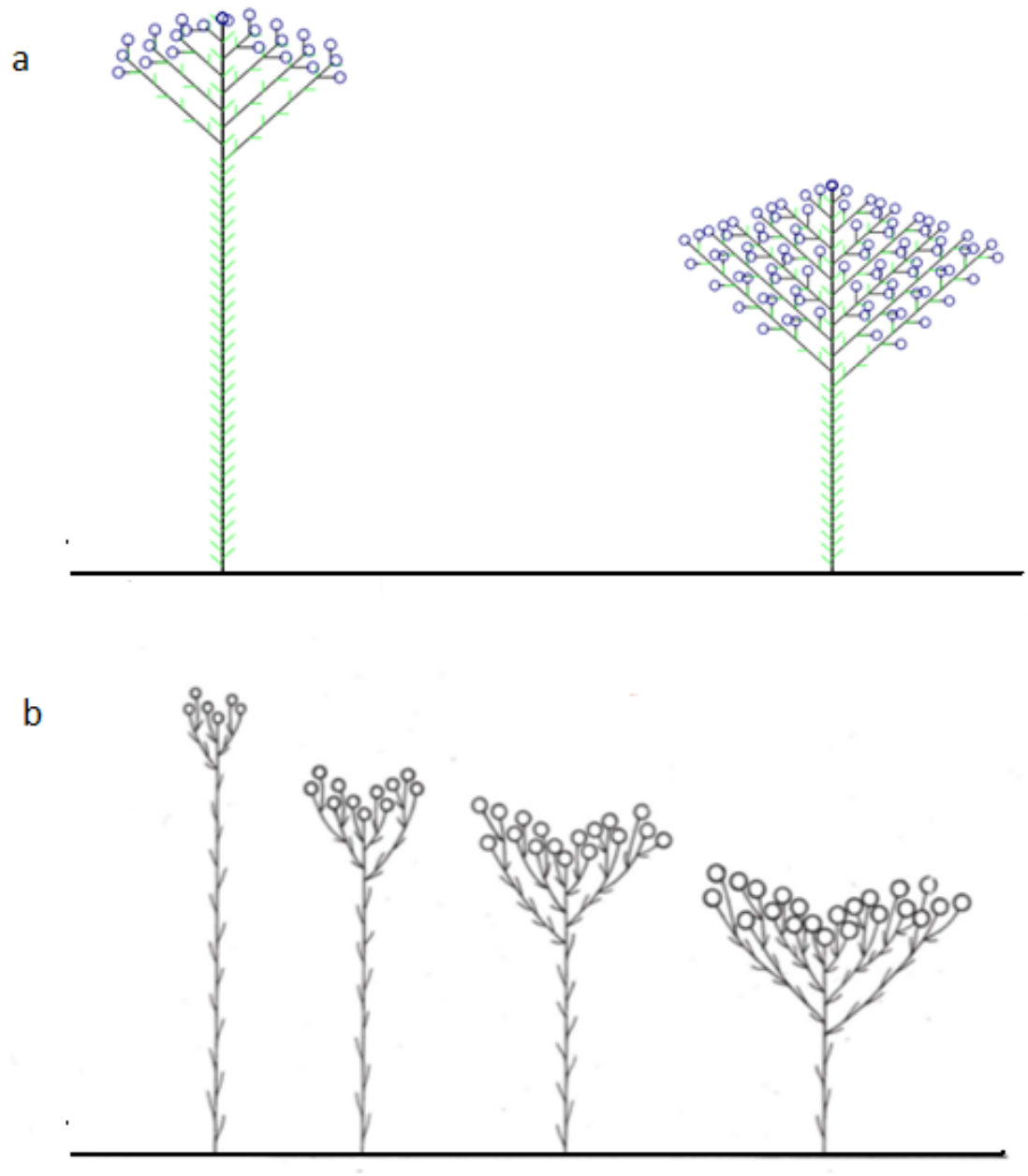


Figure 5.7: Comparison of different flax architectures. Floral architectures and plant statures of flax varieties, a) samples obtained during *PGROWTH* simulations with different *FT*/*TFL* ratios, b) flax varieties according to Kulpa and Danert (1962).

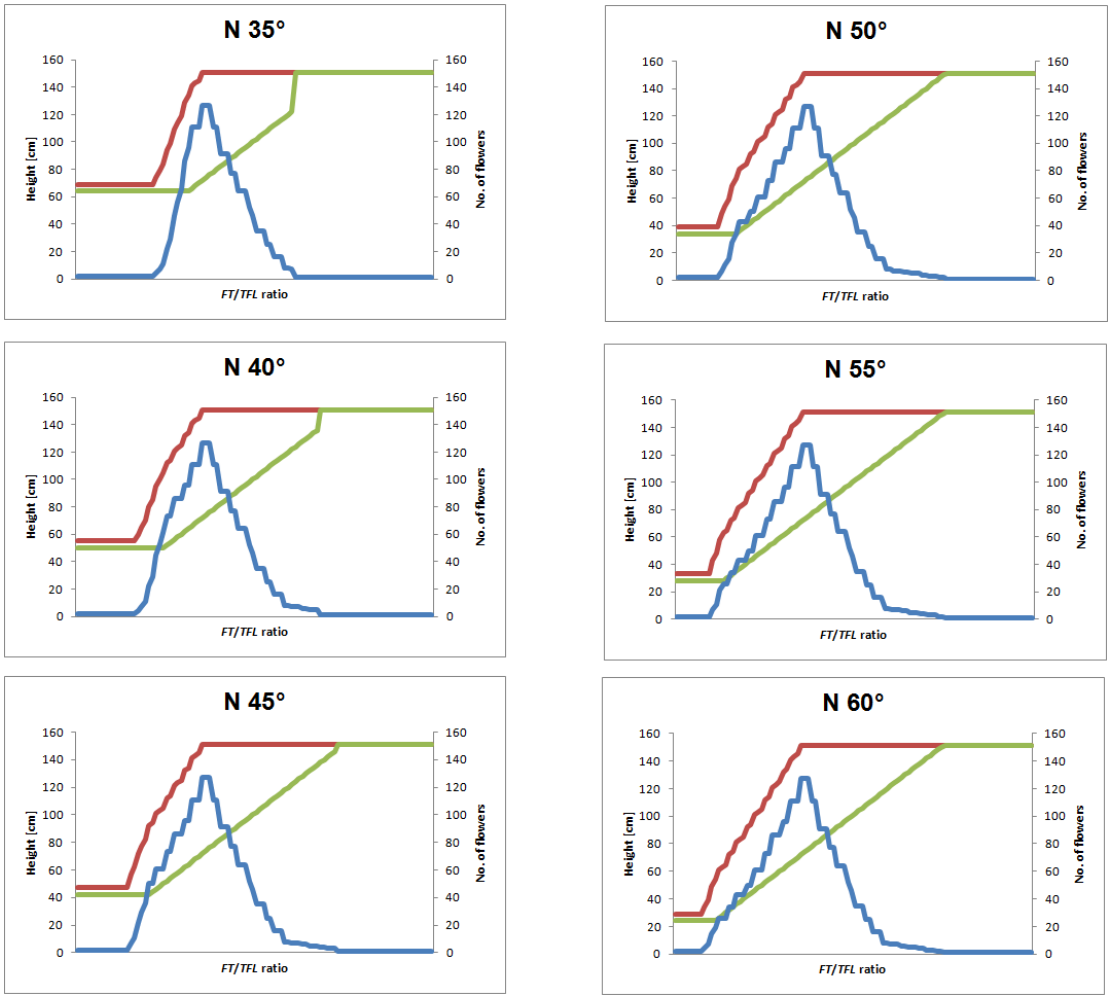


Figure 5.8: Plots showing architecture parameters obtained during *PGROWTH* simulations for various relative *FT*/*TFL* ratios on different latitudes. *FT*/*TFL* ratio increases from left to right. Blue line shows number of flowers produced, red line – total height of main stem, green line – height of main stem without floral branches.

5.4 DISCUSSION

The link between the adaptation of flax to northerly latitudes and the rise of fibre varieties is becoming more apparent in light of the results from correlation analyses and computer simulations. Time of flowering in flax is significantly positively correlated with height of the plant and negatively correlated with seed mass. Hence, the change in phenology has an impact on oil and fibre yield. This was confirmed with computer simulations, which tested the model of impact of *FT*/*TFL* on both flowering time and plant architecture. Expression of *LuFT*, *LuTFL1* and *LuTFL2* genes was measured in time-course experiment and supported their homology to *FT* and *TFL1* genes in *A*. *thaliana*. There is mounting evidence for a role of flax PEBP gene homologs in processes of floral initiation and architecture forming.

5.4.1 Correlation between flowering time and morphological traits

The results of correlation study within morphological and phenological traits in pale flax are not fully congruent with these for cultivated flax. In pale flax flowering time correlates only with plant height but not with latitude. This is against the general understanding of how phenology of plants will change across a latitudinal gradient. Furthermore, in pale flax plant height and seed mass is positively correlated, while in cultivated flax this relation is negative. This discordance might be a false result for pale flax because the dataset is based on a small sample size (only 29 samples taken). Moreover, the data was not collected for some of the samples, this is especially important in case of plant height where only half of the samples were measured. Thirdly, measurements carried out in this experiment were not congruent with measurements taken by Uysal and collaborators (2012) in the two overlapping accessions (W042 and W043). This might be due to strong environmental impact on these traits, especially on flowering time. For reliability, this experiment should be repeated with all the plant accessions in exactly the same conditions. By contrast to pale flax, the dataset used for cultivated flax is robust; altogether 1146 samples were taken into account. Hence it is assumed that associations in the data for cultivated flax are correct. In this case plant height is negatively correlated with seed mass.

Negative correlation between cultivated flax height and seed mass shows that fibre and oil varieties might have been under disruptive selection, which led to specialization in high yields of their respective products. This notion finds its support

in the two reported associations; seed size is positively correlated with oil content (Diederichsen & Raney 2006) while stem height and branching pattern are correlated with fibre content (Diederichsen & Ulrich 2009). This leads to conclusion that taller plants with smaller seeds were selected for fibre production, while shorter, big-seeded plants were favoured for oil production. Similar observation was made by Kulpa and Danert (1962) when they classified flax cultivars into specialised convarieties. It was discussed in Section ***1.3.1*** that the smaller seeds of flax and the fibre production are historically associated with flax cultivated in Central Europe and could have been consequences of adaptation to northerly latitudes. The negative correlation between seed size and plant height supports this statement and leads to conclusion that flax varieties evolved under disruptive selection.

The correlations between morphological and phenological traits are very complex. Knowledge about underlying genetic factors leads to better understanding of these associations. For example, the correlation study shows association between plant height and branching pattern and both of these are weakly associated with flowering time. It has been suggested that in *A. thaliana* expression of *TFL1* gene has an impact on flowering time, plant height and branching pattern at the same time (Prusinkiewicz *et al.* 2007) and hence it might be the case that homolog of this gene underlies similar association in flax. The early flowering model for flax includes three loci, responsible for flowering promotion, plant height and their joint regulation. This suggests a strong link between plant height and flowering time (Fieldes & Amyot 1999). However, The correlation analysis presented here indicates that plant height has a much stronger association with seed mass than with flowering time. This might suggest that smaller seeds in Central European archaeological sites (Herbig & Maier 2011) are the result of shifted resource allocation to vegetative growth through monopodial growth habit rather than phenology itself. Since *FT* and *TFL1* genes are responsible for both flowering time and determinacy of growth, plant height and seed size might be primarily associated with growth rates and growth determinacy. It has been shown that wild *A. thaliana* plants show latitudinal cline of flowering time and growth rate at the same time (Debieu *et al.* 2013). In summary, the association between vegetative growth rates and seed mass might be the primary reason for observed decrease in seed size in the northerly latitudes. This would suggest that flax adapted to the northern climate through enhanced indeterminate

growth habit. In this case, the association between flowering time and agronomic traits in flax might be secondary.

Because determinacy, flowering time and plant height are linked together through genes such as PEBP, it is likely that the adaptation to northerly latitudes could result with change in plant stature and in consequence promote one of the two purposes of cultivation. *TFL1* homologs, which have an impact on all the aforementioned traits, are perfect candidates for explaining molecular basis of this phenomenon. Since *LuTFL1* haplotypes are clearly associated with latitudes but no non-synonymous mutations are found in exons between *LuTFL1*.I and *LuTFL1*.III, expression pattern of both haplotypes was compared to screen for differences on regulatory level (Section ***5.4.2***). Link between plant height, branching pattern and flowering time was further investigated using computational modelling (see Section ***5.4.3***).

5.4.2 LuTFL1, LuTFL2 and FT expression patterns

To provide more in-depth context and improve reliability of *LuTFL1* expression analysis, additionally *LuTFL2* and *LuFT* expression levels were assessed. Further to that, the *GAPDH* housekeeping gene was used as a positive control for constant and uniform expression. In each case (including housekeeping *GAPDH*) no expression was observed in day 55 in accession W042. This is most probably due to degradation of RNA during sample processing. In other samples *GAPDH* expression is uniform confirming that it is a good target for positive control in expression studies (Huis *et al.* 2010). Further to that, the semi-qPCR results showed different expression patterns for the three genes of interest. *LuTFL2* was not expressed in any of the samples analysed here; it is either inactive in these growth conditions or expressed in other plant tissues. Alternatively, as suggested by phylogenetic analysis (Section ***3.4.3***) *LuTFL2* is related to *ATC* and as such could be up-regulated only in SD (Huang *et al.* 2012; Yoo *et al.* 2010). Because in this experiment plants were grown at 16h daylength, *ATC* homolog possibly was not expressed. In earlier studies on flax's flowering mechanism, *TFL1* homologs could not be amplified from cDNA (De Decker 2007). Probably, their target gene was a *LuTFL* homolog, which either is not expressed in targeted tissue, inactive or not expressed in the experimental conditions.

By contrast to previously described, *LuTFL1* is expressed in stems, meristems or leaves from 40 to 110 days of growth. In plants W077 and W094, uniform

expression level was observed, whereas increased expression was noted over time in other accessions. In plant W067, increased expression was observed after cold treatment period that ended in day 52. Despite the fact that on sequence level *LuTFL1* is also more similar to *ATC* than to *TFL1* (Section ***3.4.3***, Figure **3.6**), it shows the expression pattern similar to the latter. In *A. thaliana TFL1* is expressed weakly in days 2-7 but then it is up-regulated and continuously expressed in shoot meristems after inflorescence commitment (Bradley *et al.* 1997). The plasticity of functionality, protein sequence and expression pattern within PEBP homologs has been shown multiple times (Blackman *et al.* 2010; Hsu *et al.* 2011; Laurie *et al.* 2011; Pin *et al.* 2010) and hence it might be the case that in flax an *ATC* homolog evolved to be expressed like *TFL1* or alternatively *TFL1* evolved to contain one amino acid characteristic to *ATC* activity. Both *TFL1* and *ATC* function redundantly (Mimida *et al.* 2001), therefore, regardless of whether *LuTFL1* in flax is more similar to *TFL1* or *ATC* it is still likely to fulfil similar function as floral inhibitor. For clarity the putative homologs in flax, including *LuTFL1* and *LuTFL2* are named after *TFL1* rather than the *ATC* gene.

Table 5.3: Combination of *LuFT* expression data and flowering time.

Trait	W042	W043	W067	W069	W077	W094
LuFT expression in day 91						
LuFT expression in day 110						
Start of flowering [days]	87	109	103	99	100	94

In all the cDNA samples *LuTFL1* was present in two splice variants. The bulk of amplified cDNA consisted of a full gene with all four exons, however, there is also a variant without exon 3 expressed in low quantities throughout all the time points in all pale flax genotypes (Figure **5.6**). Similar, incomplete splice variant has also been observed for *TFL1* homolog in *Crocus sativus* (Tsaftaris et al. 2012). The function of the incomplete splice variant is unknown.

Finally, the expression of *LuFT* directly precedes the emergence of the first flower bud consistently in five different plant accessions. The only exception is W067,

which expressed *LuFT* very weakly in day 91, yet flowered as late as day 102. The pattern of *LuFT* expression in remaining five accessions supports the homology between *LuFT* and *A. thaliana FT* gene. The latter is also expressed just before flowering (Kardailsky *et al.* 1999). Furthermore, *LuFT* has got all the crucial amino acids that are responsible for *FT* activity, however two out of three binding pocket amino acids are different in flax than in *A. thaliana* (Section ***3.4.3***, Figure **3.5**). In this case the pattern of expression and the sequence similarity in the *FT*-activity sites congruently point to *LuFT* as structural and functional homolog of *FT*. In order to confirm the roles of *LuTFL1* and *LuFT* in flax flowering a transgenic experiment needs to be carried out. If these genes complement the *tfl1* and *ft* mutants in *A. thaliana*, it means that they are functional homologs of *TFL1* and *FT* respectively.

5.4.3 Model simulating the impact of FT and TFL1 flax homologs on flax floral architecture

Based on the sequence comparison (Section ***3.4.3***) and expression study (Section ***5.4.2***) there is a very high chance that *LuTFL1*, *LuTFL2* and *LuFT* are homologs of PEBP family and as such play integrative role in flowering. Hence, *A. thaliana* model of florigen (*FT*) and antiflorigen (*TFL1*, *ATC*), which regulate flowering through competition was accepted in *PGROWTH* script. Further to that, a model of increasing *FT* expression with vegetative growth (increasing number of stem nodes and leaves) was used in these simulations after McGarry & Ayre (2012b). Simulations covered flax growth over 150 days. As a result of simulations different flax architectures were created. The differences between plant height and floral architecture were a consequence of mutations that change theoretical *FT*/*TFL* ratio. In extreme examples simulated plant architectures resemble these that are the effect of transformations with *TFL1* and *FT* under strong promoter in tomato (McGarry & Ayre 2012b) and *A. thaliana* (Prusinkiewicz *et al.* 2007). Some of the flax architectures resembled specialized fibre and oil varieties described by Kulpa and Danert (1962). Interestingly, only these architectures that produced extremely tall plants with very high number of flowers are not observed in nature. They were probably created in *PGROWTH* due to lack of an inhibiting parameter, which should slow a proportion of vegetative or generative growth in consequence of limited resources available to plants. Apart from that, the model in *PGROWTH* script seems to be reflecting natural variation in flax architectures well.

In the second part of the experiment, *PGROWTH* model was used to investigate how different *FT*/*TFL* ratios will influence number of flowers and plant height over different latitudes. The daylength threshold parameter for flowering was set to 14 hours. This setting prevents plants that are grown at N°30 latitude from flowering at all. Further to that, the higher the latitude, the lower ratio is required for flax to start flowering, which allowed even very short plants (due to strict growth determination) to deliver progeny. This is caused by the possibility of meeting the daylength threshold earlier in the calendar with increased latitude. At the same time, the higher the latitude, the higher ratios will still allow flax to flower, which in turn renders very tall plants able to produce progeny. Similarly to the previous case, this is caused by possibility of meeting daylength threshold even later in the calendar in higher latitudes. In this way despite heavy inhibition from *TFL* (reflected by low *FT*/*TFL* ratio) flax could eventually flower late. An interesting conclusion could be drawn from this experiment, that higher latitudes allow plants to create more extreme architectures. In northerly climate flax could sustain stronger disruptive selection for fibre on the one side and oil varieties on the other. This means that even taller plants with fewer flowers could be produced in the north and is supporting our hypothesis that specialized varieties arose in Central Europe.

5.4.4 Comparing the model to the observed data

Some of the data collected during PEBP genes expression analyses could be used to validate chosen assumptions taken in *PGROWTH* model. For instance, uniform *TFL1* homolog expression was assumed during flax floral development. This finds support in the results for *LuTFL1* expression in flax. Furthermore, based on model proposed by McGarry & Ayre (2012b) the *FT* expression is not only controlled by daylength parameter but also by progress in vegetative development in flax. Again this assumption found its support in the analyses of *LuFT* gene. Despite the fact that plants were grown in cabinets with 16h of day light, the expression of *LuFT* started only in day 91 or in some cases day 110. This might reflect existence of another regulatory factor on *FT* expression such as vegetative maturity. Finally, the data for *LuFT* expression link this gene's activity with floral initiation through homology with *FT* gene in *A*. *thaliana*. There are more assumptions to be tested, such as the impact of *LuTFL* genes on flowering, however, observations made in this project give indication that *PGROWTH* model might be robust.

One of the hypotheses tested in this chapter is that plant height is correlated with flowering time in flax. Based on phenotypic observations flowering time and plant height are characterised with weak but highly significant correlation both within cultivated and pale flax. Simulations carried out based on *PGROWTH* model confirmed that such correlation exist. Plants that flower earlier in the year are shorter, later flowering plants on the other hand, are taller. The model shows that control over these two related phenotypic traits could be attributable to expression of *FT* and *TFL1* homologs.

Finally, interesting point has been made based on simulations carried out in different latitudes. It appears that plants of flax in the north could sustain stronger selection towards extreme phenotypes. More variable length of day in the European climate allows shorter and taller plants to reproduce when compared to plants that survive in the Near East. This phenomenon might have enabled selection for more specialized varieties of flax in the north, for example fibre varieties. This notion is in turn congruent with archaeological findings – fibre varieties arose in Central Europe (Maier & Schlichtherle 2011). Furthermore, decrease in seed size in northern fibre varieties could be attributed to change in resource allocation towards vegetative growth. There is a strong, negative correlation between seed mass and plant height in flax. In summary, phenotypic diversity in flax match both the output of simulations carried out with *PGROWTH* model and archaeological findings. There is no direct evidence that rise of fibre flax varieties in Central Europe was linked to adaptation to Northerly latitudes, however, all the theoretical inferences indicate that such association is indeed possible.

CHAPTER 6: THE POPULATION STRUCTURE AND MIGRATIONS WITHIN CULTIVATED AND PALE FLAX BASED ON RADSEQ MARKERS

6.1 INTRODUCTION

It is hypothesised that pale flax, which inhabits northerly latitudes of Europe, might have contributed towards the adaptation of cultivated flax introduced by early farmers to Central Europe. Genetic migration occurs between crop plants and their wild relatives. They are usually regarded as separate species but often are interfertile and therefore the genetic isolation between them is not full. For this reason they could be still considered members of one population. This population could be divided into genetically distinct subpopulations. This structure reflects isolation of subpopulations through domestication or due to geographic and climatic barriers. In flax, genetic separation is expected between wild and domesticated species. Additionally, flax subpopulations could be structured based on their geography and in case of cultivated flax alone, based on purpose of cultivation. Despite the genetic separation, gene flow is still possible between subpopulations. This phenomenon was well illustrated in the haplotype network of *LuTFL1* gene, in which genetic migration was observed between two species of flax.

The main aim of Chapter 6 is to investigate genetic structure of flax populations and measure the gene flow between wild and cultivated flax. The RADseq technique was used to develop SNP markers and genotype them across samples of cultivated and pale flax. Based on these markers, geographic structure within both flax species is investigated. It is followed by migration analyses, which will answer the questions of whether genetic migration occurred between pale and cultivated flax beyond the area of domestication and if this phenomenon was common in the evolution of cultivated flax. Additionally, genome-wide scans of loci were used to detect the signature of selection between populations of cultivated flax from different latitudes.

6.2 MATERIALS AND METHODS

6.2.1 DNA isolation and RADseq library preparation

From the total collection, 90 plants were chosen for the RADseq experiment. They represent 27 pale flax and 62 cultivated flax accessions. Plants were grown in glasshouse conditions and harvested after seedlings reached 10 cm of height. Plant material was instantly frozen in liquid nitrogen and freeze-dried in -50°C for two days. Leaves were detached from stems and weighted. Between 10 and 15 mg of dried leaves was used for DNA extraction. Material was ground with three glass beads (3 mm diameter) in TissueLyser machine (Qiagen) for 1 min with 30 Hz shaking frequency. Powdered tissue was then used for DNA isolation with DNeasy® Plant Mini Kit (Qiagen) following the manufacturer's manual. In the final step, DNA solution was collected in three subsequent elution steps, each with 50 µl EB buffer. DNA concentration was measured with Quant-it™ dsDNA Broad-Range Assay kit (Invitrogen) in Qubit® fluorometer (Invitrogen).

For RADseq library preparation, 40 µl of genomic DNA at concentration of 25 ng/µl was prepared. Further steps of library prep are described in journal protocol (Etter *et al.* 2011). In short, DNA was digested with 2U of SbfI HF restriction enzyme (New England Biolabs) for one hour and then the reaction was heat inactivated. Further to that, DNA fragments were ligated with 1 µl of custom-made, uniquely barcoded P1 adapters at 100 nM concentration (Electronic Supplement **12**) for 90 min using T4 DNA ligase, rATP, NEB 2 buffer followed by heat inactivation. Samples of ten different DNA isolates were pooled together into the total of 2 µg of DNA in 300 µl of solution. DNA was sheared ten times on ice in Bandelin Sonoplus HD 2070 sonicator with following settings: 20% power, 20% pulses for 30 s, followed by 30 s break. Sheared DNA was purified using AMPure XP beads (Agencourt) in proportion 1 to 1. Random DNA overhangs that resulted from sonication were blunted and phosphorylated using Quick Blunting Kit (NEB) followed by magnetic bead purification. Further to that, Klenow Fragment 3'-5' exo (NEB) was used to create A overhang and the library was once again bead-purified. Custom made P2 adapters were ligated to A overhang. Approximately 20 ng of library was amplified with Phusion HF Master Mix (NEB) in total of 14 cycles. Size spectrum of amplified library was visualized using HS DNA Analysis Kit in 2100 Bioanalyzer (Agilent Technologies). Further to that approximately 5 ng of amplified library was submitted

for sequencing. First 50 samples were sequenced on HiSeq 2000 Illumina platform at Oxford Genomics Centre using the TruSeq reagents. Another 40 samples were submitted for sequencing on Genome Analyzer II Illumina platform in Genomic Centre at University of Warwick. During the second run the TruSeq reagents were used.

6.2.2 Computer analyses: RADseq marker development and basic population statistics

Raw reads from Illumina sequence runs were first assembled into FastQ format and assessed in *FASTQC* software v0.10.1 (Andrews 2012) to check for standard quality measures: quality scores, overrepresented sequences, GC content and N content. Further to that, *STACKS* pipeline v1.05 (Catchen *et al.* 2011) was employed to de-multiplex and *de novo* assembly RAD sequence markers. Firstly, low quality sequences and these characterized by erroneous barcodes were discarded. Remaining sequences were sorted according to their barcodes into single-end (P1) and paired-end (P2) sequence files. Further to that, the *DENOVO_MAP.PL* script was used to call stacks of loci for all the multiplexed samples. The following settings were applied: five identical, raw reads required to create marker stack, five mismatches were allowed between alleles in single individual, 15 threads were executed and calling haplotypes from secondary reads was disabled. Marker stacks were then imported into *MYSQL* database implemented with *PERL DBD* package that allowed browsing loci with graphical interface. For loci that were characterized by very high F_{ST} values between northern and southern cultivars, paired-end reads were collated using *BOWTIE2* v2.1.0 (Langmead & Salzberg 2012) and then transformed into indexed format with *SAMTOOLS* v0.1.19 (Li *et al.* 2009). Loci that were present in all 90 samples and contained exactly one SNP were exported in *STRUCTURE* format using the *POPULATIONS.PL* script (Electronic Supplement **13**).

In order to assess the degree of differentiation between the main subpopulations of flax, F_{ST} statistic was calculated (Nei 1977; Wright 1950). For that purpose, the RADseq marker loci in structure format were transformed into a genind object using *ADEGENET R* package (Jombart & Ahmed 2011). The dataset was split between five subpopulations (three of pale flax and two of cultivated flax) and pairwise F_{ST} were measured for all the RADseq loci combined using *PAIRWISE.FST* function in *ADEGENET* and *PP.FST* in *HIERFSTAT* package (Goudet 2005). Furthermore,

heterozygosity was measured using *HS* function for the following subpopulations of flax: 1) all pale flax, 2) Turkish pale flax, 3) Greek pale flax, 4) Croatian pale flax, 5) all cultivated flax, 6) fibre varieties, 7) oil varieties, 8) intermediate varieties, 9) landraces, 10) dehiscent varieties, 11) northern cultivated flax and 12) southern cultivated flax. Additionally, for all these subpopulations a statistic similar to nucleotide diversity was calculated using *NUC.DIV* function in *PEGAS R* package (Paradis 2010; Paradis *et al.* 2004); this statistic is different from the nucleotide diversity in that it was based on the merged bi-allelic sequences rather than haplotype data. Finally, the relationships within and among pale and cultivated flax were assessed using hierarchical clustering. For this purpose the sequences of merged bi-allelic SNPs were aligned in *CLUSTALX* v2.1 (Larkin *et al.* 2007) and UPGMA algorithm employed to build a relationship tree.

6.2.3 Computer analyses: population structure of flax

A command line version of model-based clustering software *STRUCTURE* v2.2 (Falush *et al.* 2003; Pritchard *et al.* 2000) was used to make inferences about population structure in both pale and cultivated flax. Analysis was carried out in three replicates for number of populations (K) spanning between 2 and 20. Further to that, *STRUCTURE HARVESTER* (Earl & vonHoldt 2012) software was used to apply method for evaluating likelihoods of different K's (Evanno *et al.* 2005). The three replicates with the highest likelihood were summarized in *CLUMPP* software v1.1.2 (Jakobsson & Rosenberg 2007). Additionally to these analyses, script designed for inbreeding species called *INSTRUCT* v1.0 (Gao *et al.* 2007) was used to make inferences about the population structure under the assumption of high self-pollination frequency. Again, K between 2 and 20 were investigated using three independent Markov Monte-Carlo Chain (MCMC) runs. The convergence of different chains was measured using Gelman and Rubin method (1992) and together with posterior probabilities was used for choosing the best-fitted K value for this dataset. Custom-made script in *R* was used to present population structures inferred from both *STRUCTURE* and *INSTRUCT* in graphical form with desired configuration of samples.

The correlation between the latitude and RAD tag alleles within the cultivated flax was modelled using logistic regression in *R* programme for each RAD marker. Alleles in each RADseq tag were coded as binaries and correlated with centroids for

each sample (for details about how centroids were obtained see Section ***4.2.3***). Increase of allele frequency per grade of latitude was calculated using logit transformation. Subsequent to this, all the RAD tags were sorted based on p-values of correlation. The tags with the top scores were taken forward and the geographic distribution of their alleles within pale flax investigated. Additionally, F_{ST} statistic was employed to study genetic division in northern and southern populations of cultivated flax. Data for cultivated flax in genind format were transformed into loci object in *PEGAS*. Samples were segregated into two groups based on their latitude; individuals from above N40° were included in the northern subpopulation while individuals from below N40° in southern. The N40° latitude delimits Turkey and Greece from Central and Northern Balkans and is used here as approximate borderline between the distribution of domestication-associated *LuTFL1*.I haplotype and northern-associated *LuTFL1*.III haplotype in pale flax (see Section ***4.3.2*** for details). Next, the t-test was carried out for each variety to determine whether its population structure is significantly different between samples from below and above the N40° grade of latitude. Further to that, F_{ST} values were calculated between northern and southern cultivated flax population for each locus to check for signature of selection using *FST* function in *PEGAS*.

6.2.4 Computer analyses: genetic migration within and between cultivated and pale flax

Different models of genetic migration among the five flax subpopulations (three pale and two cultivated flax) were tested using *MIGRATE-N* software v3.6 (Beerli 2006; Beerli & Felsenstein 2001). Data in *STRUCTURE* format was transformed into *migrate-n* format using *PGDSPIDER* tool v2.0.5 (Lischer & Excoffier 2012). Migration models were tested in three stages: 1) migration between pale flax populations, 2) migration between cultivated flax populations and 3) combination of both including migration events between the two species. In the first stage, two scenarios were tested a) migration occurring only between adjacent subpopulations to account for geographic barriers and b) unrestricted migration between all the subpopulations that reflect their glacial refugium ancestry. The two respective scenarios were also explored in cultivated flax, in this case however, the reasoning behind unrestricted migration model is that humans spread flax in Eurasia through continuous trade. In stage three, the best-fit model was applied for migration within

species. Furthermore, the two scenarios of interspecies relationship were explored: a) migration from pale flax occurred only in the Near East, reflecting the gene flow in the area of domestication and b) migration from pale flax occurred multiple times in different geographic regions, which would imply post-domestication gene flow to cultivated flax. All the models were ranked with marginal likelihood estimated with Bayesian inference and MCM chains approach. In the search for optimal parameters one long chain and four short, heated chains (with 1.0, 1.5, 3.0, 1000000 heating scheme) were used. The number of steps for analysis was set to 5000, by the end of estimation 1000000 of the first samples were discarded. Prior thetas were set as uniform and spanned between 0 and 10 while prior migration rates - from 0 to 2000. These intervals were inferred from multiple, preliminary *MIGRATE-N* searches and additionally were suggested in the literature for similar analysis (Hubner *et al.* 2012). Due to high number of cultivated flax (34 samples per population) in comparison to pale flax (nine samples per population), a subset of twelve individuals per population was randomly chosen for each loci.

Additionally to migrate-n analyses, a maximum likelihood approach implemented in *TREEMIX* software v1.1 (Pickrell & Pritchard 2012) was used to infer about possible migrations in flax populations. A custom made script was written in the R programming language to help in transforming data into *TREEMIX* format. All of 219 RADseq markers were used in estimating a tree of the seven populations and possible migration events between them. No SNP grouping was option was used during the analysis as none of the loci are assumed to be influenced by linkage disequilibrium. Between one and three migration events were explored during ML tree estimation, the higher limit is dictated by minimum number of loci required for such analyses. The final graph was visualized in *R* using the *PLOT_TREE* function, which is included in the *TREEMIX* package.

6.3 RESULTS

6.3.1 RADseq marker development and basic population statistics

As a result of data processing with the *STACKS* pipeline a catalogue of unique RADseq markers (so called RAD tags) was created in a *MYSQL* database. Within 90 individually barcoded samples a total of 359775 tags were created. The number of RAD tags that were present in all 90 samples is 1263. The average sequence coverage per individual for the P1 tags ranged from 30 to 422. Within these tags that were present in all the individuals 607 were polymorphic, while 219 contained only one SNP. Only the tags with single SNP were used in further analyses.

Table 6.1: Pairwise F_{ST} values between cultivated and pale flax subpopulations based on RADseq data.

	PF Turkey	PF Greece	PF Croatia	CF South	CF North
PF Turkey	-				
PF Greece	0.0810	-			
PF Croatia	0.0902	0.0415	-		
CF South	0.2123	0.2408	0.2437	-	
CF North	0.2652	0.2831	0.2990	0.0233	-

Table 6.2: Heterozygosity and modified bi-allelic nucleotide diversity of all flax subpopulations and varieties based on RADseq data.

Group	Heterozygosity	Nucleotide diversity	No. of samples
Pale flax	0.2165	0.0864	28
Turkish	0.2140	0.0887	10
Greek	0.1974	0.0678	11
Croatian	0.1894	0.0641	7
Cultivated flax	0.2424	0.0978	62
Fibre varieties	0.2225	0.0792	23
Oil varieties	0.2386	0.0941	20
Intermediate	0.2162	0.0713	8
Landraces	0.2103	0.0730	7
Dehiscent	0.2113	0.0818	4
North	0.2293	0.0792	37
South	0.2536	0.1190	25

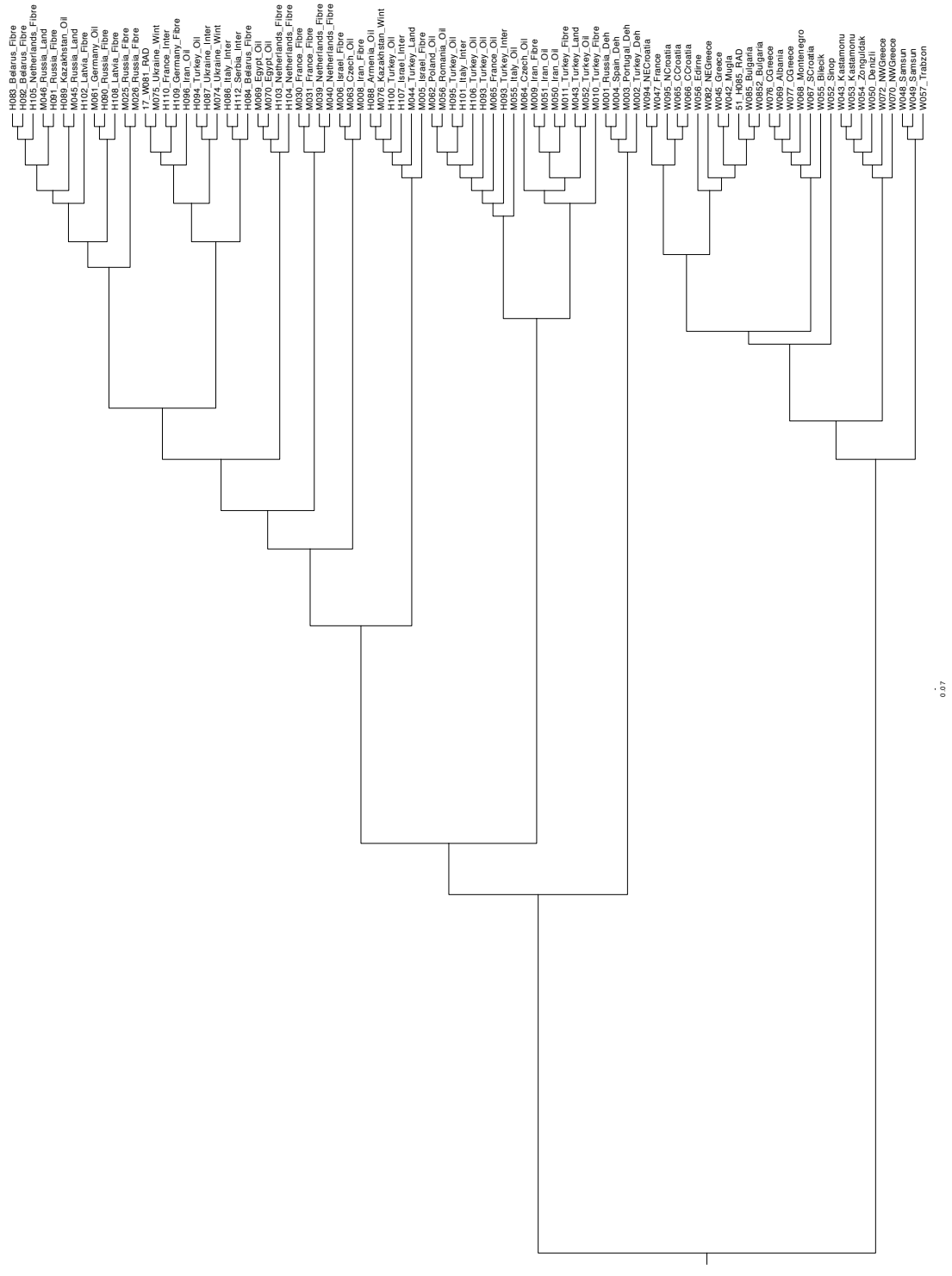


Figure 6.1: Cladogram based on UPGMA clustering of RADseq bi-allelic data for 90 samples of flax. The labels consist of accession names (where W – wild, H – historic cultivated and M – modern cultivated flax), country of origin and, in case of cultivated flax only, variety.

The genetic separation within the five main subpopulations (three within pale flax and two for southern and northern cultivated flax respectively) was measured (Table **6.1**). Southern and northern cultivated flax populations were very similar to each other (F_{ST} = 0.02). The relationships among pale flax populations are more complex, with Turkish accessions being more separated from European populations (F_{ST} = 0.08 and 0.09) than the two European between themselves (F_{ST} = 0.04). The genetic division between southern cultivated flax and pale flax populations is slightly lower (F_{ST} ranging from 0.21 to 0.24) than between northern cultivated flax and pale flax (F_{ST} from 0.26 to 0.30).

Bi-allelic nucleotide diversity and heterozygosity was greater in cultivated than in pale flax (Table **6.2**). Within cultivated flax, the southern population is characterized by greater diversity than northern population and at the same time oil varieties are more diverse than any other groups, including landraces. The cladogram based on sequences of merged bi-allelic SNPs was rooted with pale flax clade (Figure **6.1**). Turkish samples from the eastern-most region were placed in the most basal position, followed by other Turkish accessions. The more derived clade includes two sister groups: the one of which contains south-western pale flax samples and the second, which consists of northern samples from Croatia, Bulgaria and France together with samples from east Balkans. Within cultivated flax the most basal position was taken by dehiscent varieties. Following to that, there are two sister clades; the first groups accessions mainly from Turkey and Iran, from which great majority are oil varieties. The second clade contains six groups: five of them contain mixed accessions, while the most derived one consists predominantly of north-eastern fibre varieties.

6.3.2 Population structure in flax

Flax population is structured primarily based on the division between pale and cultivate flax. Additionally, dehiscent varieties could be easily separated. In the unrooted phylogram based of F_{ST} values dehiscent varieties are stemming from between pale and cultivated flax (Figure **6.2**). Applying different colours to different geographic locations of analysed samples allowed structuring them spatially. Even though there is some overlap, pale flax from Turkey (green) is clustered together while accessions from the Balkans (orange) are grouped separately. Geographic pattern for cultivated flax is less clear, however, enrichment of southern varieties

(red) is visible closer to dehiscent flax, while distal branches are predominantly northern (olive green).

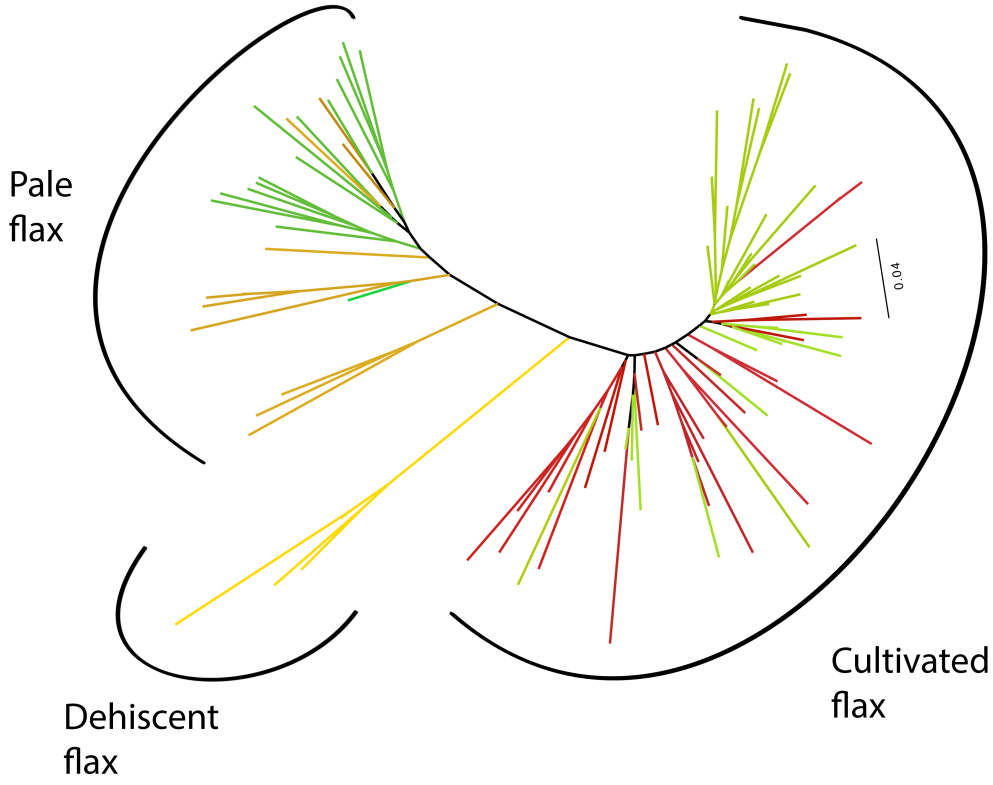


Figure 6.2: Unrooted phylogram based on UPGMA clustering of RADseq data. In pale flax clade samples from Turkey are marked in orange, from the Balkans – in green. Dehiscent varieties of flax are marked with yellow colour. Within flax cultivars the Near Eastern varieties are marked in red, while European in green.

In the *STRUCTURE* analyses the highest delta K parameter was calculated for number of populations K = 6 (Supplement **2**). Graphs for less likely K are presented in the supplement (Electronic Supplement **14**). Within three replicates of analyses for K = 6 no major differences were found, hence results were merged using the *CLUMPP* script (Figure **6.2b**). Individuals were arranged in the major groups: pale flax, dehiscent varieties, landraces, intermediate, fibre and oil varieties. Within these groups they were sorted by their latitude from south to north. The great majority of Turkish pale flax accessions (nine out of ten) and one Greek sample belong to subpopulation I (orange colour, Figure **6.3b**), two Greek samples from the West Coast to subpopulation II (purple), while the rest of pale flax populations to subpopulation III (green). One of the aforementioned Greek samples and sample from Western Turkey have got admixture of subpopulation IV (yellow), which is the main component of

all four dehiscent flax varieties. The cluster with oil varieties predominantly belong to subpopulation V (red) with various levels of admixture from I, II, III and IV. Finally, the other varieties belong mainly to subpopulation V (red) in the southern latitudes and to VI (olive green) with higher latitudes. The divergence between all the subpopulations is summarized in Table **6.3**.

Table 6.3: Allele frequency divergence among subpopulations in *STRUCTURE* analysis.

	Sub I	Sub II	Sub III	Sub IV	Sub V	Sub VI
Sub I orange	-	0.0301	0.0355	0.1051	0.0645	0.0863
Sub II purple		-	0.0331	0.0485	0.0202	0.0338
Sub III green			-	0.0968	0.0744	0.0945
Sub IV yellow				-	0.0968	0.0944
Sub V red					-	0.0277
Sub VI olive						-

In *INSTRUCT*, the dynamics of change of Deviance Information Criterion was used to establish that number of populations K = 4 best fits the data under the assumption of self-fertilization (Supplement **3**). Graphs for other K parameters are included in the supplement (Electronic Supplement **15**). The convergence of the three separate MCM chains was calculated with Gelman-Rubin statistics to 1.058. In the *INSTRUCT* graph for K = 4, only eastern-most samples from Turkey belong to subpopulation I (red), which they share with the cultivated flax (Figure **6.3a**). The two western Greek pale flaxes have admixture from subpopulation III (yellow), which in turn is predominant in dehiscent varieties. The rest of pale flax populations are assigned to subpopulation II (green). The admixture from this subpopulation is visible in small amounts in cultivated flax, mainly in oil varieties but also in southern samples of other varieties. The northern samples of the other varieties predominantly belong to subpopulation IV (olive green). Again, individuals were arranged in major groups: pale flax, dehiscent varieties, landraces, intermediate, fibre and oil varieties and within - sorted by increasing latitude. The results of t-test determined that the genetic structure of oil varieties is not significantly different between samples from below and above the N40° grade of latitude (p-value = 0.32) while the genetic structure of remaining varieties is (p-value = 1.65E-06). The graph for cultivated flax is similar between the *STRUCTURE* and *INSTRUCT* analyses.

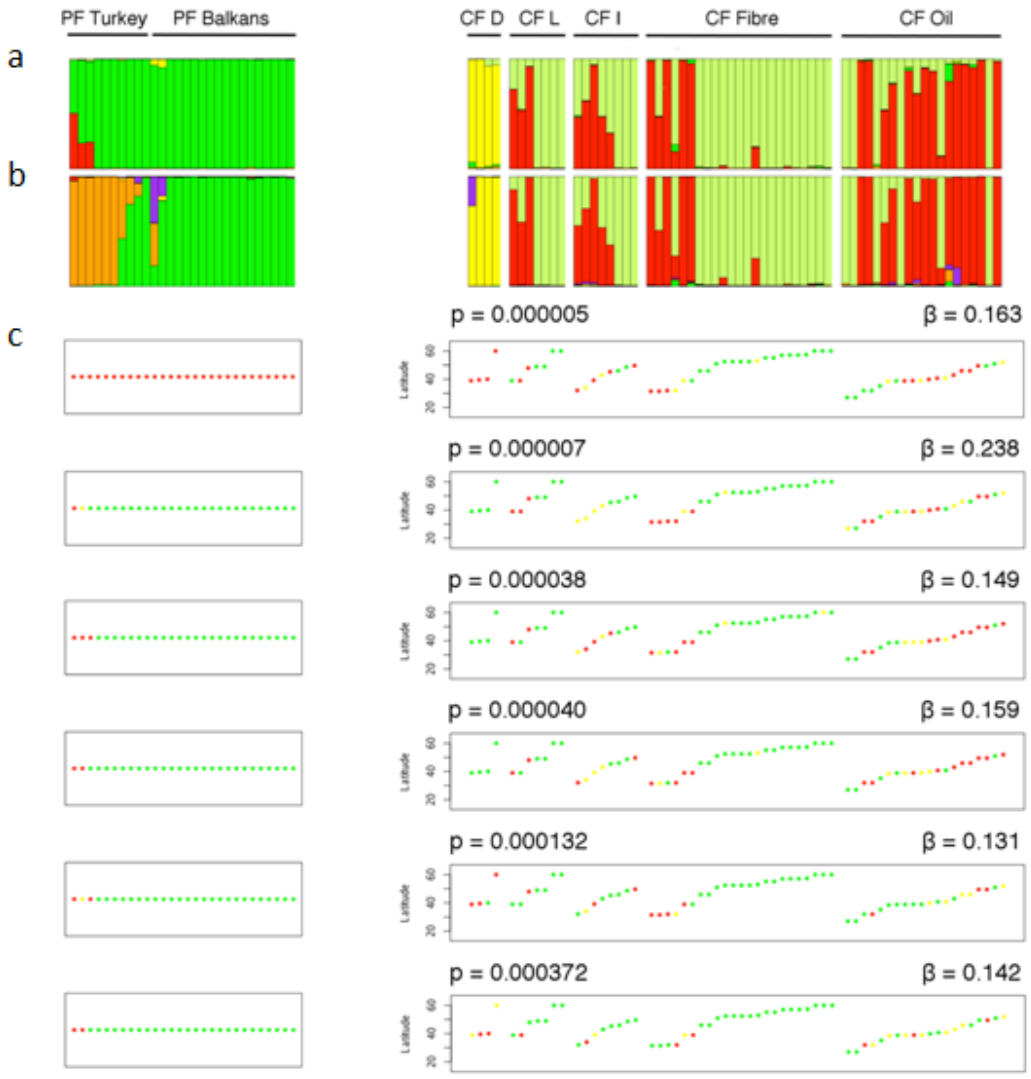


Figure 6.3: Population structure of pale flax and varieties of cultivated flax based on RADseq data. The abbreviations used above the figure denote: PF – pale flax, CF – cultivated flax, CF D – dehiscent varieties, CF L – landraces, CF I – intermediate varieties. Within these groups samples were arranged according to increasing latitude.
a – population structure based on *INSTRUCT* analysis
b – population structure based on *STRUCTURE* analysis
c – the detailed distribution of alleles in six RAD tags with the strongest correlation with the latitude.

To see which loci contribute the most towards the north-south division in landraces, intermediate and fibre varieties of cultivated flax the logistic regression was calculated over latitudes for each locus. The most significantly correlated six loci were characterized by p-value lower than 0.001. Interestingly, five of these loci are also polymorphic in pale flax with a clear phylogeographic pattern (Figure **6.3c**). In

all of this five RAD tags the ancestral allele is associated with pale flax accessions from the eastern-most Turkey and southern cultivated flax at the same time. The derived allele on the other hand, is present in remaining pale flax populations and in northern cultivated flax.

The contribution of each RADseq locus to genetic division of northern and southern cultivated flax was additionally measured using F_{ST} scores.. In this analysis, F_{ST} values between northern and southern populations of cultivated flax for all the RAD tags were calculated and arranged in the increasing order. All the six RAD tags, which correlated strongly with latitude, were highlighted with red colour in the F_{ST} analysis (Figure **6.4**). The two tags with the highest F_{ST} values overlap with tags that contributed the most towards latitudinal cline in population structure of landraces, intermediate and fibre varieties.

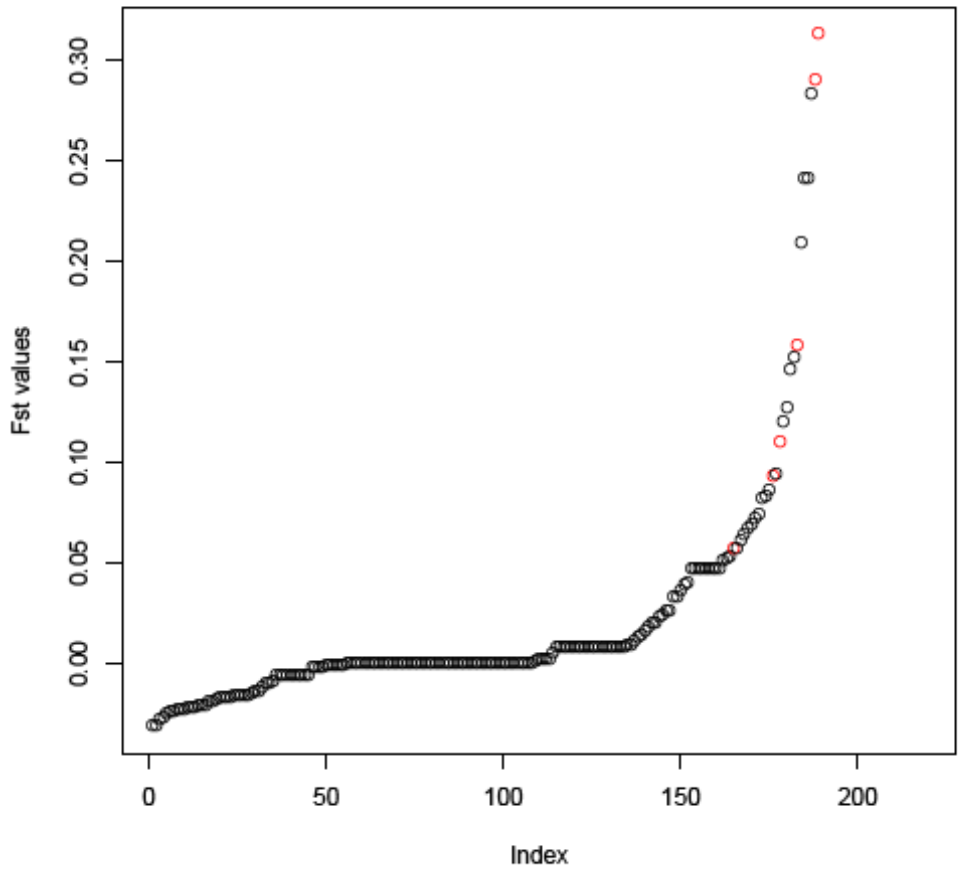


Figure 6.4: F_{ST} values for RADseq tags dividing northern and southern populations of cultivated flax. In red are the tags that contributed the most towards latitudinal cline in population structure of landraces, intermediate and fibre varieties.

6.3.3 Genetic migration within and between cultivated and pale flax

MIGRATE-N analyses revealed that a full-migration model best explains RADseq allele frequencies within three pale and two cultivated flax populations. Models with restricted migration from pale to cultivated flax, and vice versa were characterized by lower marginal likelihood values (Table **6.4**). The population mutation rates (Θ) and effective number of immigrants per generation (N_{em}) were calculated for each population. Both estimators are similar across all the populations: Θ varies from 0.039 to 0.048, while N_{em} ranges from 5.27 to 7.77. A map summarizing all the effective migration rates per population is presented in Figure **6.5** together with F_{ST} values between them.

Table 6.4: Marginal likelihoods of migration models tested in *MIGRATE-N* analyses.

Populations	Model	Bezier score
Pale flax	Full migration	13755.07
	Geographic restrictions	10317.84
Cultivated flax	Full migration	22874.83
	Geographic restriction	12681.53
Combined	Full migration	38774.17
	Gene flow during domestication only	21761.28

Maximum likelihood tree estimation with migration events in *TREEMIX* resulted with only one significant migration (Figure **6.6**). When two and three migration events were tested the tree topology was not consistent across ten replicates. The tree itself consists of four cultivated flax populations that are nested within three pale flax populations. The most ancestral population in cultivated flax comes from the Near East, while the most derived population – from Russian region. The most closely related to cultivated flax is the pale flax population from Turkey. Greek and Croatian populations are more distant from cultivated flax, however, the clearest evidence for migration occurred from between Greek and Turkish pale flaxes to cultivated flax from southern, Mediterranean region.

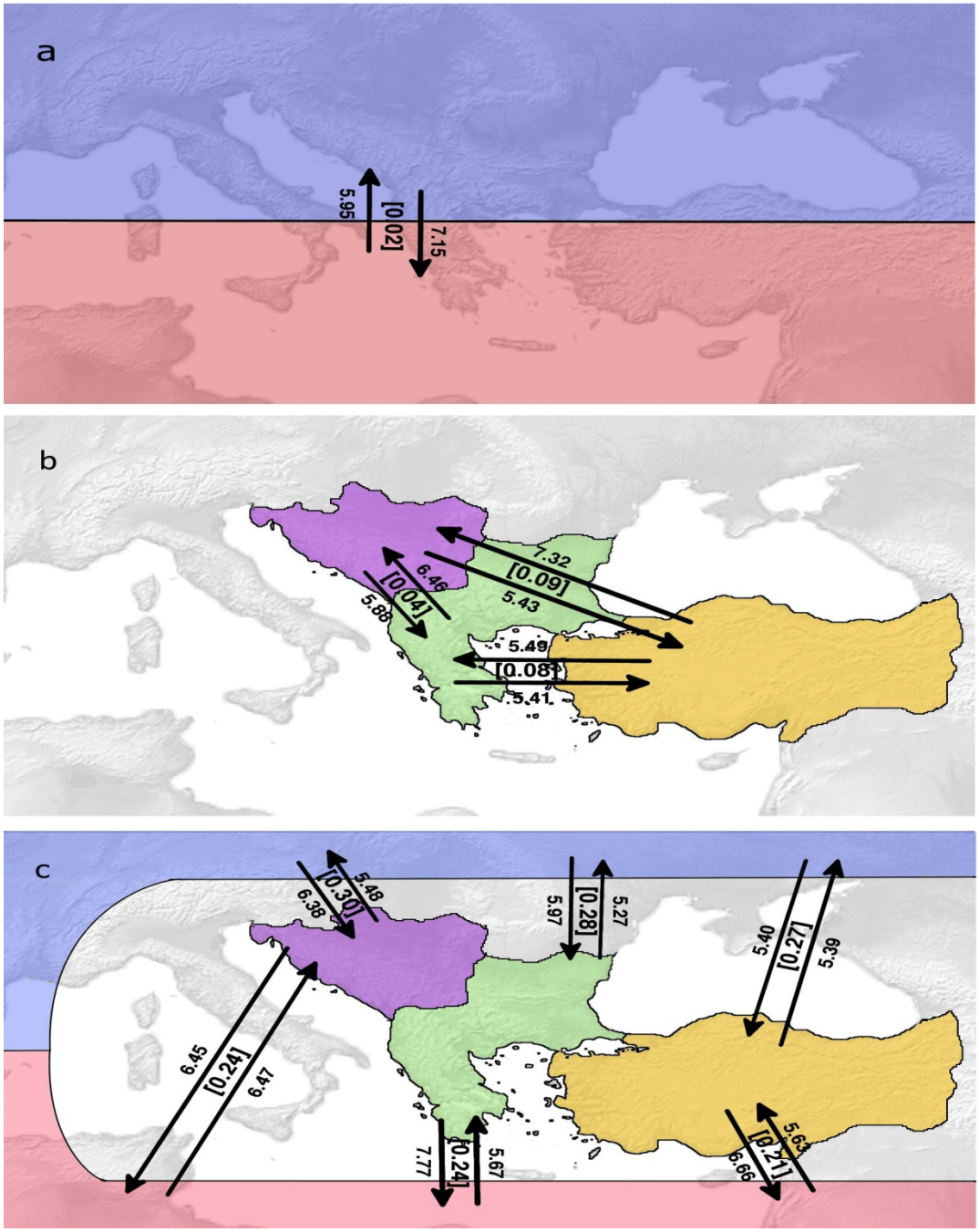


Figure 6.5: Maps with marked distributions of flax subpopulations and migration events. Arrows between flax subpopulations indicate migration; the outer numbers represent the effective number of immigrants per generation, inner number in square brackets represent F_{ST} values.
a – genetic migration between southern (< 40°N, in red) and northern (> 40°N, in blue) populations of cultivated flax
b – genetic migration within three European pale flax subpopulations (purple, green, orange)
c – genetic migration between pale and cultivated flax.

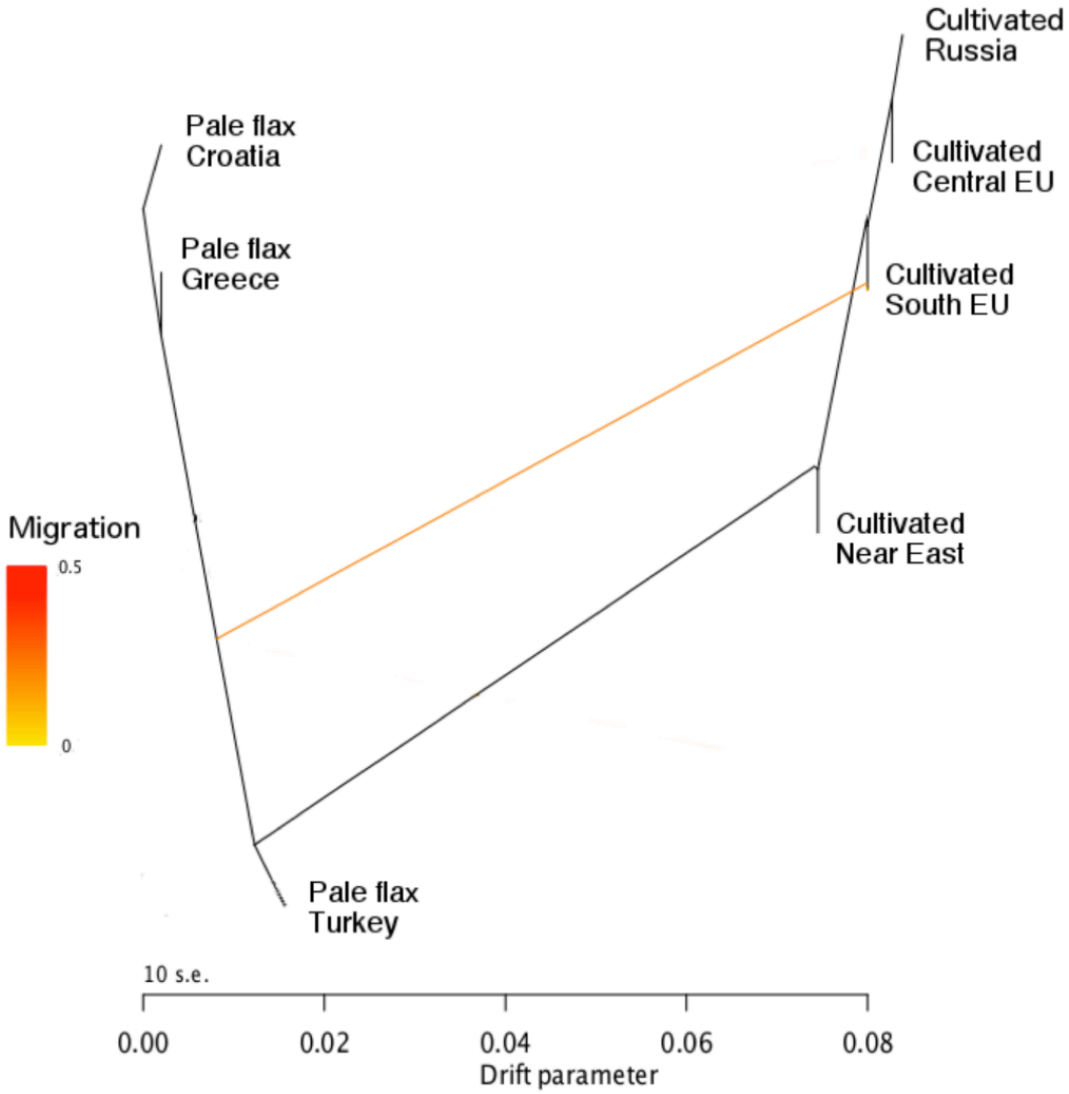


Figure 6.6: *TREEMIX* analysis of three pale and four cultivated flax populations. Orange line indicates major migration event between pale and cultivated flax.

6.4 DISCUSSION

The RADseq approach for marker development and genotyping in pale and cultivated flax was very successful. Despite relatively low cost and labour required for this type of studies, it enabled discovery of 219 polymorphic loci within studied 90 samples of flax. All of these markers are characterised with only one SNP; the remaining RADtags were discarded to avoid confusion that could arise during assignment of polymorphic characters to two alleles. The number of markers and genotyped samples in this study is sufficient for phylogeographic approach, population structure inference, clustering and basic population genetics analyses. In this section, firstly the genetic diversity of both species of flax is discussed. The results presented here support literature reports about very small reduction of genetic diversity during the domestication. Furthermore, the genetic structure of flax populations is investigated. There are interesting patterns of admixture between populations and species of flax. That leads to investigation of gene flow between populations and species of flax. The hybridization scenario is discussed and evidence shown for increased frequency of wild alleles from beyond the domestication area in northern populations of cultivated flax.

6.4.1 The genetic diversity of pale and cultivated flax

Differences in the genetic diversity of pale and cultivated flax populations can be helpful in understanding the evolution history of both species. Genetic variation in cultivated flax is expected to be lower when compared to its wild progenitor, because only a subset of wild species' genetic diversity is captured during the domestication. This is caused by an expectation of a relatively small size of founding population involved in the process. The effective population size of wild rice and maize that explains genetic diversity of modern cultivars could have been very small (>1500 individuals) but diverse (Eyre-Walker *et al.* 1998; Zhu *et al.* 2007). As a consequence of small founding population we usually observe around 30% drop of genetic diversity in most crop plants (Gepts 2004). The effect of narrowing the genetic diversity is referred to as domestication bottleneck.

Cultivated flax also seemingly suffered from domestication bottleneck and heavy inbreeding. In recent years the breeding community was alarmed by reports of very low morphological and genetic diversity especially within fibre flax varieties

(Everaert *et al.* 2001; Vromans 2006). Contrary to these reports, data presented in this thesis suggest existence of relatively high genetic diversity in cultivated flax even in fibre varieties. Fibre varieties are characterised by a lower heterozygosity than oil varieties, but higher than intermediate varieties, landraces and pale flax (Table **6.2**). In general, heterozygosity and bi-allelic nucleotide diversity within cultivated flax is higher than in pale flax. Bi-allelic nucleotide diversity is not a standard measure but allows comparing diversity between populations by employing the original formula (Nei & Li 1979) to sequences of merged pairs of SNPs. The conclusion that cultivated flax has got higher genetic diversity than its wild relative is supported by other studies. Analyses of flax genomic DNA have shown that there are significantly more exclusive SNPs in cultivated flax (162) then in pale flax (3) (Fu & Peterson 2012), while based on ISSR and EST markers, genetic diversity between the two species is similar (Fu 2011; Uysal *et al.* 2010). This might indicate that the impact of domestication bottleneck on cultivated flax was weak. Coalescent simulations support this notion and indicate that initial reduction of diversity reached no more than 27% (Fu 2012). In case of the singular-event domestication, population bottleneck is expected to be much more severe. Hence, the small reduction of flax diversity could be a strong evidence for diffused-origin domestication of this crop plant. Further to that, the genetic variation within cultivated flax could have increased due to either rapid evolution, post-domestication gene flow with its wild relatives in the vast area of their coexistence or both factors simultaneously. In order to assess which of these factors played the most important role in cultivated flax evolution, one has to investigate the geographic structure and population admixture in both species.

6.4.2 The population structure and admixture levels of pale and cultivated flax

The pale flax population structure revealed in this project could be explained with isolation by distance model and presence of geographic barriers. Balkan subpopulations are more similar to each other than to Turkish pale flax based on F_{ST} values. Genetic exchange between the latter was narrowed by presence of the seas between Greece and Turkey. A dendrogram based on bi-allelic sequences show similar pattern of clustering: the most of Turkish pale flax accessions form distinct clade (Figure **6.1**). This additionally is supported by *STRUCTURE* analysis, in which predominantly Turkish samples were marked with orange colour (Figure **6.2**). By

contrast, based on *INSTRUCT* analysis, which takes into account inbreeding habit in flax, only three eastern-most Turkish samples were separated from within pale flax cluster (red colour). Finally, there is a monophyletic clade for northern-most pale flax samples from Croatia and France in the dendrogram, however, this group was not partitioned out in structure analyses. Possibly the signal of population structure between Greek and Croatian pale flax was blurred by hybridization events in refugium area during last glacier advance.

Within cultivated flax, genetic divergence between southern and northern populations is very low based on F_{ST} values. The differentiation between them is not very clear in the dendrogram. However, the most derived clade of the dendrogram is significantly enriched with fibre varieties from North-eastern Europe. Similarly, the most basal clade of indehiscent flax is significantly enriched with oil varieties from the Near East. In *STRUCTURE* and *INSTRUCT* graphs, the division between northern and southern samples is very clear within landraces, intermediate and fibre varieties but not in oil varieties. Southern samples are predominantly belong to subpopulation I (red) while northern to IV (olive green). This is supported by the results of t-test, which showed that the structure separation with N40° latitude is significant in any other varieties than oil. The lack of latitudinal genetic structure in oil varieties and presence of such structure in other varieties could be explained in two ways. Firstly, it could be an effect of undisrupted gene flow between oil varieties and latitudinal isolation between other varieties. Secondly, it could reflect relatively late adaptation of oil varieties to northern latitudes and in consequence non-diverged population structure. This reasoning is congruent with archaeological finds that report decreasing flax seed size over time in Neolithic Central Europe (Herbig & Maier 2011) that is interpreted in context of early adaptation of fibre varieties (Maier & Schlichtherle 2011) and difficulties in adaptation of oil varieties.

The levels of admixture between pale and cultivated flax are generally low. The *INSTRUCT* analysis indicates a close relationship between eastern Turkish pale flax (W048, W049 and W057) and cultivated flax, which could reflect a domestication event. Turkish pale flax clearly contributes towards the genetic make-up of oil varieties and southern samples of the other cultivated flax varieties. In the dendrogram samples W048, W049 and W057 are in an ancestral position and within pale flax are the most closely related to cultivated flax. Geographically, they are the

closest to the Near Eastern centre of domestication and hence their genetic similarity to cultivated flax supports the notion that flax was primarily domesticated in western Anatolia (Helbaek 1959; Zohary & Hopf 2000). UPGMA analysis clustered all the Turkish pale flax samples together with the two samples from North-eastern Greece (W070 and W072). These two Greek samples are admixed with dehiscent varieties in both *STRUCTURE* and *INSTRUCT* analyses. This indicates that pale flax populations from Greece might have contributed towards emergence of these early varieties (Fu 2012; Fu & Peterson 2010). Furthermore, the similarity between Greek pale flax and cultivated flax is only slightly lower than similarity between Turkish pale flax and cultivated flax based on F_{ST} values (Table **6.1**). This is especially clear in their respective relationship with northern cultivated flax and might suggest that the genetic impact of Greek pale flax on cultivated flax was as great as the genetic contribution from wild populations of Western Anatolia during domestication. There is strong evidence for admixture in structure analyses that confirms impact of the Balkans pale flax on dehiscent, fibre and oil varieties of cultivated flax. This result calls for more in-depth analysis of gene flow between the both species.

6.4.3 Migration and post-domestication gene flow between the two flax species

Pairwise migration indexes were estimated in *MIGRATE-N* programme. Usually a higher number of individuals per population than is used here for flax is required for migration analysis. In the case of the dataset used for flax, there are only 90 samples, however, there are 219 SNP markers analysed, which should secure a good resolution of this study. Furthermore, all of the parameters estimated by *MIGRATE-N* have narrow peaks of posterior distribution, suggesting robust statistical basis for this estimation (Electronic Supplement **16**). Finally, the results were consistent for all three datasets tested (pale, cultivated and mixed flax), which further supports the robustness of this analysis. In the first two stages of analysis models of migration were chosen for populations of pale and cultivated flax separately. The results of the first test suggest that migration occurred among all pale flax populations, despite the fact that Croatian and Turkish subpopulations are not adjacent. This is could be due to gene flow that occurred during the glacial period, when populations from northern Balkans had to retreat to southern refugium (Barton & Hewitt 1985). This scenario should be reflected by high diversity in Greek subpopulation. Interestingly though, Turkish pale flax is genetically more diverse than the Greek pale flax. Secondly, the

all-migration model is the most likely for cultivated flax populations as well, suggesting that anthropogenic migration may have nullified geographic distance.

Models of migration in the combined analysis (including both pale and cultivated flax) were tested and all-migration model once again was characterized by the highest marginal likelihood (Table **6.4**). This supports the notion that post-domestication gene flow from pale to cultivated flax and *vice versa* occurred multiple times in their evolutionary history. Interestingly, migration rates per generation are similar on both intra- and inter-species level (Figure **6.5**). It indicates that gene flow within pale flax populations is as strong as gene flow between pale and cultivated flax, despite high differences in F_{ST} values. Interestingly, these results are different from the results of similar experiment in another inbreeding crop species – barley. Although, the migration rates within wild barley populations are very similar to these observed for pale flax, the migration rates from wild to cultivated species are much lower in barley when compared to flax (Hubner *et al.* 2012). The migration rates within pale flax populations are similar to these between pale and cultivated flax. This is surprising in that one would expect higher migration rates within species than between species. However, the two flaxes diverged not longer than 12,000 years ago and it is likely that isolation barriers are still very weak.

Cultivated flax spread throughout the Near East and Europe at a very quick pace due to human migrations and trade, finally reaching beyond the geographic distribution of its wild relative. The intensive gene flow between cultivated and pale flax in the area where they overlap could be attributed to: 1) rapid introduction of foreign, interfertile species into direct neighbourhood of local populations and/or 2) seed admixture during early agricultural traditions, which might have been supplemented by foraging. There is robust evidence for coexistence of farmers and foragers in Central Europe (Bollongino *et al.* 2013) and the Balkans (Boric & Price 2013). Additionally gene flow from pale flax might have had an adaptive value for cultivated flax and hence more immigrant alleles were fixed in its population. The analysis of loci that contribute significantly towards latitudinal gradient in cultivated flax revealed that in five out of six instances the northern allele originated from other pale flax populations than these in the domestication area (Figure **6.3c**). The allele that was inherited from pale flax during domestication was replaced in northern cultivated flax by the allele from other pale flax populations. The analyses of genetic

division between northern and southern cultivated flax populations suggest that these six RADseq tags are characterized with very high F_{ST} values and hence could have been under selection (Figure **6.4**). In summary, it might be the case that the genetic migration from pale to cultivated flax might have been inflated by intake of adaptive alleles.

In this chapter evidence is presented for gene flow from Greek and Croatian pale flax to both northern and southern cultivated flax. *TREEMIX* analysis supports occurrence of migration between Greek-Turkish pale flax and southern cultivated flax (Figure **6.6**). At the same time, *MIGRATE-N* analysis resulted with the highest effective migration per generation (N_{em} = 7.77) from Greek pale flax to southern cultivated flax. This suggests that the highest influx of genes from wild to cultivated flax occurred around Bosphorus strait. Furthermore, *STRUCTURE* analysis revealed low levels of admixture from Balkans' pale flax to both southern and northern cultivated flax (Figure **6.3b**). According to the results obtained from *MIGRATE-N*, gene flow occurred from all pale flax populations to northern cultivated flax, however, in lower rates than towards southern populations. This is additionally supported by F_{ST} values, which indicate lower divergence between wild populations and southern flax than wild populations and northern cultivated flax (Table **6.1**). Together, the results of these analyses suggest higher frequency of gene flow from pale flax populations to southern cultivated flax but support existence of such gene flow to northern varieties as well.

The lower, visible migration rates from pale flax to the northern populations of cultivated flax could be caused by the fact that the coexistence of pale flax and southern varieties has a longer history than the coexistence of pale flax and northern varieties that emerged later (Herbig & Maier 2011). It could be additionally influenced by the fact that big proportion of northern flax is cultivated outside the distribution of wild relative. The gene flow between pale flax and northern cultivated flax occurred historically. Despite the fact that pale flax does not grow in Russia, Russian fibre varieties have admixture from pale flax. Therefore, the gene flow from pale flax to northern cultivated flax could be an effect of historic migrations, which occurred before an adaptation of cultivated flax to Northern Europe.

In conclusion, there is a strong evidence for post-domestication gene flow between pale and cultivated flax that possibly occurred during the spread of early farmers into Europe. Based on the neutral RADseq markers this gene flow was the strongest between pale flax from around Bosphorus strait towards southern European cultivated flax. Furthermore, there is strong evidence, that within loci that were adopted from pale flax outside the domestication area, there are at least five that are associated with cultivated flax growing in the northerly latitudes. This recapitulates the *LuTFL1* scenario. The *LuTFL1* haplotype I inherited by cultivated flax during the domestication was replaced in northern populations by haplotype III, which originated in the middle Balkans pale flax. This line of inference is further discussed in **Chapter 7**.

CHAPTER 7: GENERAL DISCUSSION

7.1 CULTIVATED FLAX WAS DOMESTICATED IN PROXIMITY OF THE FERTILE CRESCENT

Based on archaeological evidence, flax cultivation started together with other founder crops (Zohary & Hopf 2000) in the Near East about 12 000 years ago (Hillman 1975; Van Zeist & Bakker-Heeres 1975). There is a robust evidence in molecular data that cultivated flax was domesticated from pale flax (Allaby *et al.* 2005; Fu & Allaby 2010; Fu *et al.* 2012; Gill & Yermanos 1967) however, no marker dataset has yet been precise enough to establish the region of its geographic origin. In this study, the RADseq dataset analysed with *INSTRUCT* and UPGMA clustering pointed eastern Turkey, as the closest area to flax domestication. In the *INSTRUCT* analysis the southern cultivated flax have high levels of admixture with pale flax from around Turkish cities of Trabzon and Samsun (Section ***6.3.2***, Figure **6.3a** – colour red). In an UPGMA analysis these three samples of pale flax are placed in the basal position respective to cultivated flax (Section ***6.3.1***, Figure **6.1**). Both results support the origin of cultivated flax in Eastern Anatolia and is congruent with the results obtained by investigating *LuTFL1* haplotype network. This network suggests the gene flow from pale flax populations from eastern and southern Turkey to the southern cultivated flax in haplotype I, which is interpreted as associated with the domestication event (Section ***4.3.2***, Figure **4.1a**). This evidence further supports the Eastern Anatolian origin of cultivated flax, which is congruent with general point of view about the origin of European agriculture.

7.2 CULTIVATED FLAX SPREAD FROM THE NEAR EAST INTO CENTRAL EUROPE

It is possible to predict the spread of cultivated flax towards Northern Europe by dating and mapping the Neolithic archaeological remains of flax. Beyond the Near East, flax appeared first in Greece around 6000 years BCE (Valamoti 2011) and then it was recorded in Alpine region of Central Europe continuously from 4000 years BCE (Herbig & Maier 2011; Maier & Schlichtherle 2011). To the best of the author's knowledge, until now there has not yet been any genetic evidence that would support or refute this scenario.

Estimates of the genetic distances between cultivated and pale flax populations using RADseq data in *TREEMIX* software suggest that Turkish pale flax is a direct ancestor to the Near Eastern populations of cultivated flax (Section ***6.3.3***, Figure **6.6**). The cultivated flax lineage leads from the Near Eastern to South-European populations, followed subsequently by Central- and East-European populations. The order of branching reflects the direction of geographic spread of cultivated flax. Based on the data collected in this project it is difficult to predict the time of divergence of cultivated flax populations using the molecular clock approach. This method requires full-length sequences and therefore SNP data is insufficient. Furthermore, there is not a sufficient number of samples in this project to test demographic models with IMa2 software. Hence, the presented analyses cannot predict the precise time of the introduction of cultivated flax to Europe. They do however, support the notion that cultivated flax was first grown in the Near East and was later introduced to Southern Europe, followed by Central and Eastern Europe.

The geographic structure of cultivated flax populations is congruent with its genetic structure. In *STRUCTURE* and *INSTRUCT* graphs there is visible latitudinal gradient in population structure within landraces, dehiscent, intermediate and fibre varieties (Section ***6.3.2***, Figure **6.3**). A similar latitudinal gradient is observed for haplotype data based on *LuTFL1* and *LuTFL2* phylogeographic analysis. Based on logistic regression, the haplotype clusters I and III of *LuTFL1* network are associated with southern and northern latitudes respectively (Section ***4.3.2***, Figure **4.1b**). The data support the notion that cultivated flax was domesticated in the Near East (Figure **7.1**) which implies the northern populations are descendants of southern populations. Such an interpretation of data support south-to-north direction of cultivated flax spread.

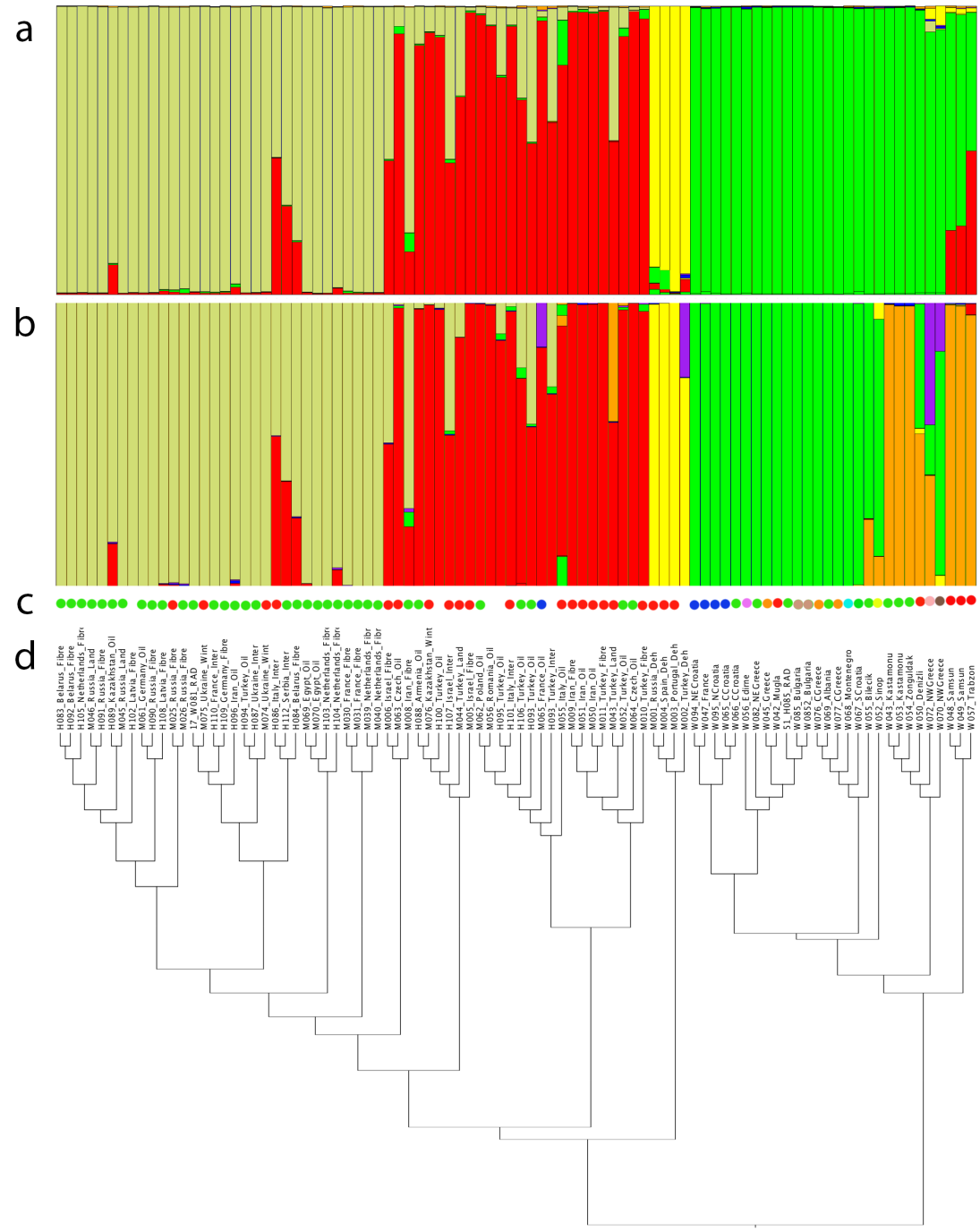


Figure 7.1: **The graph integrating the population structure and clustering analysis of flax based on RADseq data with mapped *LuTFL1* haplotypes**.
a – the population structure based on the *INSTRUCT* analysis
b – the population structure based on the *STRUCTURE* analysis
c – map of *LuTFL1* haplotypes, where: red – haplotype I, orange - II, green – III, pink – IV, teal – V, dark brown – VI, light brown – VII, blue – VIII, purple – IX and yellow – X.
d – the UPGMA clustering analysis, where the labels consist of accession names (W – wild, H – historic cultivated and M – modern cultivated flax), country of origin and, in case of cultivated flax only - variety.

7.3 POSSIBLE ROLE OF *LuTFL1* IN THE ADAPTATION OF CULTIVATED FLAX TO THE NORTHERLY LATITUDES

Plants introduced to new geographic regions may have to adapt to respective climatic conditions. By contrast to the Near East, the European climate is characterized by more variable daylength, severe cold winters and is less predictable in general. These conditions cause difficulties in the cultivation of the Near Eastern varieties of crop plants in Europe. This is especially relevant to flax, as it is one of the latest-maturing traditional crops. The climate in the northern zone increases the risk of frosts before seed maturation (Dribnenki 2010). Flax grown in the north, similarly to other crop plants, is usually insensitive to vernalization (exhibit spring growth habit) and daylength that might lead to later flowering and indeterminate growth habit (Darapuneni *et al.* 2014). The above-mentioned changes in flax phenology were probably a result of successful adaptation to the northerly latitudes.

Both the vernalization- and photoperiod-sensitivity pathways are coupled in a single gene network. The overall signal of flowering is integrated through the genes of PEBP family. The expression of these genes have a simultaneous impact on both flowering time and plant stature in *A. thaliana* (Bradley *et al.* 1997; Prusinkiewicz *et al.* 2007). PEBP gene family groups flowering promoters (such as *FT*) and inhibitors (such as *TFL1*). In this thesis, evidence is presented that *LuTFL1* is in fact a *ATC*/*TFL1* homolog. Firstly, there is a remarkable similarity between the two loci on the gene arrangement (Section ***3.3.3***, Figure **3.2**) and sequence (Section ***3.3.4***, Figure **3.4**) levels. Furthermore, all the active centre amino acid are conserved in *LuTFL1* when compared to *ATC* gene (Section ***3.4.3***, Figure **3.5**). Finally, a time course experiment showed that *LuTFL1* is expressed in distal parts of flax stems throughout both vegetative and generative growth phase, that resembles the expected *TFL1* expression pattern (Section ***5.2.2***, Figures **5.3**, **5.4** and **5.5**).

The *TFL1* homologs were involved in the adaptation to northerly latitudes in sunflower, barley and soybean (Blackman *et al.* 2010; Comadran *et al.* 2012; Tian *et al.* 2010) and might have played the same role in the evolution of flax. Similar to its homologs in other crop plants the *LuTFL1* haplotypes in flax show association with latitudes (Section ***4.3.2***, Figure **4.1**b). Haplotype I is probably a fully functional gene inherited by cultivated flax during the domestication. Haplotype XII is characterized by non-synonymous mutation, which is close to protein's active centre when tertiary

structure is taken into account. It might have had an impact on binding pocket conformation and hence reduced protein's inhibiting properties. This haplotype is common within European winter varieties and is enriched with northern samples of cultivated flax. Furthermore, haplotype III is the most frequent within northern samples of cultivated flax. However, mutations in this sequence are located in gene introns and thus are difficult to analyse for their impact on protein expression and/or activity. The analyses of genetic diversity within cultivated flax revealed that *LuTFL1* evolved under selection or a selective sweep (Section ***4.3.3***, Table **4.3**). Despite the fact that the exact mechanism of adaptation to northerly latitudes remains unknown, this study provides evidence that *LuTFl1*.III was under positive selection in the north and probably played a role in flax adaptation to European climate through change in phenology and growth determinacy.

7.4 CULTIVATED FLAX HAS BEEN UNDER DISRUPTIVE SELECTION

Flax is an unusual crop plant due to double purpose of its cultivation. On one hand, it is grown for its oil-rich seed, on the other - for its fibrous stems. Oil content is positively correlated with seed size (Diederichsen & Raney 2006) while fibre content is correlated with stem height and branching pattern (Diederichsen & Ulrich 2009). The correlation analysis carried out in this project for over a thousand cultivated flax samples shows that seed size and plant height are negatively correlated (Section ***5.3.1***, Figure **5.2**) that supports the expectation that both agronomic traits cannot be selected for at the same time. Plants with smaller seeds were selected for fibre production, while shorter, big-seeded plants were favoured for oil production. A similar observation was made by Kulpa and Danert (1962) when they classified flax cultivars into specialised convarieties. Furthermore, this line of evidence is congruent with the output of simulations carried out in the programme *PGROWTH* (Section ***5.3.3***, Figure **5.7**). Fibre and oil yield improvement require the opposite resource allocation during plant development, in vegetative or generative growth respectively. Hence, it is likely that the two varieties have been under disruptive selection.

One way of identifying disruptive selection in flax is to investigate its population structure based on molecular data. Strong structure could support the presence of such selection. RADseq data used in this project allowed the discernment of four and six subpopulations in *INSTRUCT* and *STRUCTURE* analyses, respectively (Section ***6.3.2***, Figure **6.3**). However, none of these were consistent with the division between fibre and oil convarieties of flax. Contrary to the expectations, landraces, intermediate, fibre and oil varieties had a similar genetic make-up. The difference was only observed after sorting the samples within varieties according to the latitude. It turned out that landraces, intermediate and fibre varieties are characterized by strong division between northern and southern samples. Oil varieties on the other hand, do not have such structure. Even though the separation of oil and fibre varieties is not possible based on the RADseq molecular markers in this study, an important difference was observed: fibre flax was subject to a disruptive selection in different latitudes while oil flax was not.

The signature of disruptive selection for southern- and northern-adapted varieties of flax might reflect the impact of adaptive haplotypes on genetic isolation. The geographic structure of flax populations was also noticed during the investigation of

LuTFL1 haplotype data. There is a correlation between haplotype variants and the latitude (Section ***4.3.2***, Figure **4.1b**). The genetic network of *LuTFL1* haplotypes marks a clear separation between southern cluster I and northern cluster III (Section ***4.3.2***, Figure **4.1a**). Finally, it was shown that the distribution of haplotypes in *LuTFL1* network could be explained by the presence of selection or a selective sweep (Section ***4.3.3***, Table **4.3**). In light of this evidence, it becomes clear that latitude might have had a huge impact on genetic separation of northern flax populations not only through isolation by distance but by the means of selection. This notion further supports the presence of disruptive selection in cultivated flax. However, this selection is likely to be driven by the need to adapt to northerly latitudes through change in flowering habit.

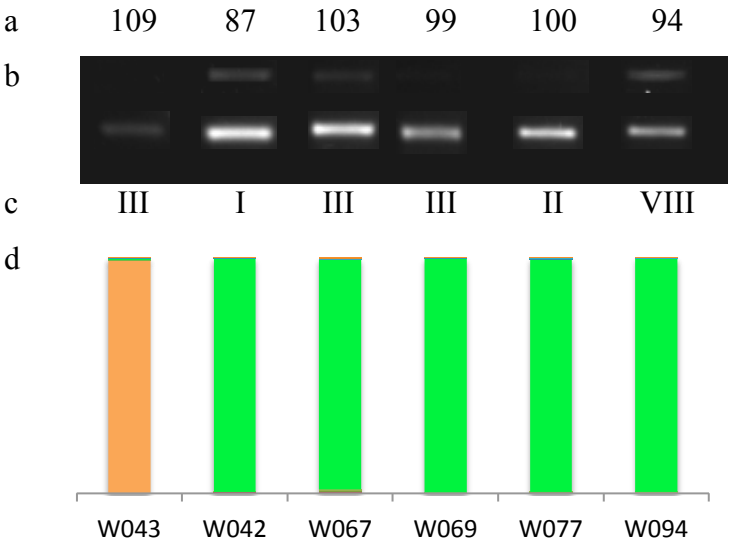


Figure 7.2: Flowering time in six accessions of pale flax and their genetic background.
a – Time from sowing to the occurrence of first floral buds in days.
b – Strength of expression of *LuFT* on 91st and 110th day after sowing.
c – LuTFL1 haplotype.
d – Assignation to genetic populations as measured by *STRUCTURE*.

There might be other genetic factors that alter flax flowering strategies. The genetic network of flowering in model plants is very complex (Flowers et al. 2009; Hall et al. 2011) and there are likely many genes regulating this process in flax. Flowering time measured in six wild accessions of pale flax was highly variable even between two

Turkish accessions. Hence, some diversity in this phonological trait existed prior to adaptation to Central European climate. Flowering time diversity is independent from variation within *LuTFL1* gene since it is not correlated with its haplotypes (Figure 7.2). The number of tested samples is far to small to make any statistically significant inferences, however, these results are congruent with expression pattern of *LuFT* gene. Additionally, late flowering accession W043 is genetically separated based on *STRUCTURE* analyses and is the only accession from the direct neighbourhood of the domestication centre. It might be the case that some other genetic factor played a role in changing cultivated flax flowering time through regulation of *LuFT* gene even before it left Turkey.

LuTFL1 might have conferred flax adaptation to Central European climate through change in mode of growth (from determinate to indeterminate) rather than change in flowering time. Change from determinate to indeterminate growth was crucial in the adaptation of soybean to Northern China (Tian et al, 2010). A remaining question pertinent to this study is whether the flax adaptation to Central Europe was linked with the emergence of fibre varieties.

7.5 THEORETICAL INFERENCE ABOUT THE EMERGENCE OF FLAX FIBRE VARIETIES

Tracing the origins of fibre flax would be much easier if variety-specific molecular markers were developed. Discrimination between intermediate, oil and fibre varieties was unsuccessful based on markers such as the *SAD2* locus (Allaby *et al.* 2005) and chloroplast loci (Fu & Allaby 2010), followed by RAPD (Diederichsen & Fu 2006; Everaert *et al.* 2001), ISSR (Uysal *et al.* 2010), IRAP (Smykal *et al.* 2011) and EST analyses (Fu 2011). The only success in this matter was a result of an AFLP study, but even then only PCA analysis in contrast to phylogenetic approach enabled such discrimination (Vromans 2006). Molecular markers specific to fibre varieties are scarce. The RADseq markers developed in this project are moderately successful in distinguishing oil and fibre varieties. Only three out of 219 RADtags are characterized by F_{ST} values higher than 0.20. At the time of writing the list of genomic regions that might be under disruptive selection between fibre and oil flax has been published (Soto-Cerda *et al.* 2013) and could be effective in investigating the origin of flax fibre varieties. For the purpose of this thesis however, another line of inference was taken.

Archaeological evidence suggests that fibre varieties of flax emerged in Central Europe during the Neolithic. The primary purpose of flax cultivation was probably for its oil rather than for fibre (Allaby *et al.* 2005). Early farmers who settled in Central Europe brought with them winter oil varieties, which were later replaced with summer fibre varieties (Diederichsen & Hammer 1995). This hypothesis is congruent with Central European archaeological findings, which record a shift in size of the flax seeds from large (associated with oil production) in lower strata to smaller (associated with fibre production) in the higher ones (Herbig & Maier 2011). Further to that, archaeologists noticed an increasing number of technological and agrarian innovations leading to more effective textile production between 4000 and 2500 years BCE in Central Europe (Maier & Schlichtherle 2011). Despite the strong support for this scenario in the archaeological data, there is little known about the biological processes that led to fibre flax emergence.

It is possible to make inferences about the evolution of flax based on some archaeological findings. The process in which seed size is decreasing in time is against the domestication trend; it is speculated that seed size increases in early stages of crop evolution (Purugganan & Fuller 2009). Hence, it is likely that the

evolution of flax in the Central Europe represent the stage of crops local adaptation. There is a natural trend in *A. thaliana* to produce smaller seeds with increasing latitude as a consequence of adaption to northerly climate (Li *et al.* 1998). The local adaptation stage of crop evolution often leaves a signature of disruptive selection in the molecular diversity of a population. It was discussed in Section ***7.4*** that such signature was found in RADseq and *LuTFL1* data and reflects the genetic dissociation of northern and southern flax cultivars. Similarly, in barley it is anticipated that population structure is correlated with climate adaptations (Jones *et al.* 2011). Furthermore, based on phenotypic data signature of disruptive selection would be expected in the division between fibre and oil varieties in population structure, this however, was not observed in the presented data.

There is a link between the adaptation to the northerly latitudes and change in yield traits of fibre flax. Adaptation to a northern climate usually involves change in flowering time and/or determinacy. Fibre yield traits on the other hand depend on stem height and branching pattern. Flax flowering time and plant height is controlled by the same set of genes (Fieldes & Amyot 1999). In this thesis, the *PGROWTH* model presents the evidence that the change of flowering time and growth determinacy has a huge impact on branching pattern and stem height (Section ***5.3.3***, Figure **5.8**). The model assumes that the shift from vegetative to generative growth stage is a major determinant of plant architecture. It models the expression of two regulatory genes *TFL1* and *FT* and their impact on timing of this shift. Assumptions of this model were taken from the observations made in tomato (Lifschitz & Eshed 2006; Lifschitz *et al.* 2006; McGarry & Ayre 2012b) and *A. thaliana* (Prusinkiewicz *et al.* 2007), however, flax inflorescence development was taken into account. Multiple simulations based on the *PGROWTH* model led to conclusion that indeterminacy causes flax plants to grow taller with fewer floral branches. Northern grown Canadian flax is characterized by indeterminate growth habit and in consequence its stems enter senescence later (Dribnenki 2010). Green flax stalks deliver more flexible fibres, which might be better for textile production. It may be the case that oil varieties of flax which arrived from the Near East changed their growth habit to indeterminate, enabling better survival at higher latitudes but compromising resource allocation to the generative phase. Consequently, the adaptation of domesticated

varieties to northerly latitudes may have compromised oil-producing abilities, but serendipitously improved fibre production resulting in a shift of use of the crop.

7.6 GENE FLOW BETWEEN PALE AND CULTIVATED FLAX WAS COMMON IN THE EVOLUTION OF THIS DOMESTIC CROP.

There is evidence that inter-species admixture has affected population structure, which might suggest that hybridization has occurred between pale and cultivated flax outside the domestication area. The population structure analyses based on RADseq data show admixture from the Balkans pale flax populations to cultivated flax (Section ***6.3.2***, Figure **6.3**). The most notable is the admixture from the two eastern Greek samples towards cultivated flax, indicating that pale flax of the region might have contributed towards genetic pool of dehiscent varieties. Further to that, we observe other cases of admixture from pale flax outside the domestication area, especially towards fibre and oil varieties. Despite the fact that pale flax does not grow in Central and Northern Europe, cultivated flax individuals from these regions have admixture from pale flax. Therefore, the gene flow from pale flax to northern cultivated flax could have been an effect of historic migrations, which occurred before an adaptation of cultivated flax to Northern Europe.

The model of panmictic population was chosen in the *MIGRATE-N* analysis for the pale and cultivated flax together. The models with restricted gene flow within and between pale and cultivated flax were characterized with lower marginal likelihoods (Section ***6.3.3***, Table **6.4**). This supports the notion that post-domestication gene flow from pale to cultivated flax and *vice versa* occurred multiple times in their evolutionary history. Interestingly, migration rates per generation are similar on both intra- and inter-species level (Section ***6.5.3***, Figure **6.5**). It indicates that gene flow within pale flax populations is as strong as gene flow between pale and cultivated flax, despite high differences in F_{ST} values (Section ***6.3.1***, Table **6.1**). This is surprising in that one would expect higher migration rates within species than between species. However, the two flax species diverged no longer than 12,000 years ago and it is likely that their isolation barriers are still weak. The gene flow between pale and cultivated flax is biologically possible as it was shown that they cross-pollinate at low rates (Gurbuz 1999) and that both species are interfertile (Muravenko *et al.* 2003; Yermanos & Gill 1969).

The post-domestication gene flow from pale to cultivated flax might have played a major role in adaptation to local conditions. In order to test that, the loci that contributed the most towards the latitudinal gradient of population structure in

landraces, intermediate and fibre flax were investigated (Section ***6.3.2***, Figure **6.3c**). Within six such loci the origins of northern-associated alleles were investigated. In five instances the northern allele was derived from pale flax populations beyond the domestication area. This illustrates the post-domestication gene flow events from wild to cultivated flax. Additionally, the frequencies of alleles in cultivated flax suggest that the inflow alleles were favoured in the northern climate. This is a similar case to that observed for *LuTFL1* haplotypes. The haplotype III, which is associated with northerly latitudes and was under selection or selective sweep, was inherited from pale flax populations inhabiting both sides of Bosphorus strait (Section ***4.3.2***, Figure **4.1c**). Since, according to all evidence, *LuTFL1* is a homolog of *TFL1* (Section ***3.4.3***) and the latter was shown to play a role in adaptation to northerly latitudes, it is concluded that the transfer of haplotype III from pale to cultivated flax might have led to adaptation to European climate.

Pale flax alleles have had an increased chance of fixation in cultivated flax populations due to their adaptive properties. In neutrally evolving populations the immigrant alleles would be often purified by the genetic drift. If they were under selection however, their frequency would increase and eventually they would be fixed in a population. This phenomenon might have caused an elevated number of immigrants per generation from pale to cultivated flax estimated by *MIGRATE-N* programme. The results of the *TREEMIX* analysis suggest that the main migration occurred from pale flax to southern populations of cultivated flax (Section ***6.3.3***, Figure **6.6**). Such gene flow must have increased genetic diversity of southern cultivars. This is congruent with Vavilov's (1951) observations; he placed the Mediterranean region as one of the main centres of cultivated flax phenotypic variation. However, the influx of the potentially adaptive alleles associated with the northerly latitudes occurred from pale to northern cultivated flax. In effect, the strength of migration estimated by the *MIGRATE-N* analysis is similar towards southern and northern cultivars (Section ***6.3.3***, Figure **6.5**). It seems that influx of northern-associated alleles might have determined the genetic structure in landraces, intermediate and fibre varieties of flax possibly due to disruptive selection between northern and southern flax. Interestingly, it did not have an impact on genetic structure of oil varieties. It is hypothesized that genetic influx from pale flax is weak

towards oil varieties, because it would introduce unwanted traits such as dehiscence and seed mass reduction.

The long-standing adaptations with complex molecular mechanisms that are inherited from pale flax might have an advantage over simple, blunt mutations that evolved within cultivated flax. Pale flax evolved for at least hundreds of thousands of years and inhabited northerly latitudes during glacier retreats. This allowed pale flax to develop sophisticated adaptive mechanisms with limited trade-offs. By contrast, the cultivated flax evolved for the last 12,000 years and therefore had to develop adaptations quickly. The easiest way to do it is through simple, crude mutations, which might bring with them deleterious effects or little adaptive flexibility (Allaby *et al.* 2014). For instance, it has been shown that some mutations that lead to change in protein structure might decrease protein's further evolvability (Tokuriki & Tawfik 2009). Within *LuTFL1* network there are two haplotypes that are associated with northerly latitudes: III, which originated within pale flax gene pool and XII, which emerged within cultivated flax (Section ***4.3.2***, Figure **4.1a**). Mutations in haplotype III occurred in introns and their putative role remains elusive. The non-synonymous substitution in exon of haplotype XII however, might have had an impact on protein structure and in consequence its activity. Interestingly, the data presented here suggest that haplotype III have been preferred over haplotype XII in the north. This demonstrates that in fact the long-standing, sophisticated adaptations might have an advantage over simple, crude adaptations.

7.7 DIVERSITY OF FLAX

Because of the population bottleneck associated with the domestication, genetic diversity of pale flax is expected to be greater then that of cultivated flax. A warning was reported that the latter was low especially within fibre flax varieties (Everaert *et al.* 2001; Vromans 2006). More recently published papers however, indicate that flax cultivars are as genetically diverse as their wild relatives. There are significantly more exclusive SNPs in cultivated flax (162) then in pale flax (3) (Fu & Peterson 2012). Based on ISSR (Uysal *et al.* 2010) and EST markers (Fu 2011) the levels of diversity in both species are similar. The sequencing data for 24 loci showed that the reduction in the genetic diversity during the domestication was not greater than 27% (Fu 2012). Only based on IRAP marker data pale flax is significantly more diverse then its cultivated equivalent (Smykal *et al.* 2011), however, these results could be largely biased by low number of pale flax accessions (11) used in this study compared to cultivated flax (708). Based on 219 RADseq markers developed and genotyped in this study the heterozygosity and molecular diversity within pale and cultivated flax are similar. Interestingly, the genetic variation is similar between landraces, intermediate and fibre varieties. The latter especially were reported to suffer from genetic erosion, however, based on above evidence their genetic diversity is similar to other varieties and landraces of flax. The evidence suggest that levels of diversity are similar in both pale and cultivated flax. It might mean that cultivated flax recovered from the founding population effect by either rapid evolution and/or frequent gene flow from pale flax or alternatively that the domestication bottleneck has had very little impact on genetic diversity of flax.

7.8 ADDRESSING THE HYPOTHESIS

The hypothesis posed in this thesis is that:

Northern-adapted, fibre varieties of flax have had significant input from northern latitude wild flax populations in support of the scenario that the evolution of fibre use is associated with northern gene flow.

The alternative scenario is that:

Northern-adapted, fibre varieties of flax are a subset of the diversity of domesticated oil producing flax and are the product of selective breeding within that gene pool in support of the notion that the evolution of fibre use was independent to wild gene flow in cultivated flax.

This study confirmed that the post-domestication gene flow occurred between pale flax populations inhabiting area beyond the Near East and the cultivated flax. Furthermore, the evidence is presented that some of the immigrant alleles are associated with the northern cultivated flax and might have contributed towards the adaptation to European climate. Finally, the theoretical basis for understanding how such adaptation might have influenced the flax architecture and in consequence result with emergence of fibre varieties are discussed.

7.9 FUTURE WORK

This study has expanded our understanding of the evolution of cultivated flax and the emergence of fibre varieties. However, further validation is necessary to link the processes of adaptation to northerly latitudes and the emergence of fibre flax. The most suitable approach would be to test if haplotypes of loci such as *LuTFL1* have an impact on plant architecture and yield traits. For this purpose, transgenic experiments should be carried out. In order to observe the effect on plant architecture one could transform *A. thaliana tfl1* mutants with *LuTFL1* haplotypes. In the next step, knockout *lutfl1* mutants could be created in flax followed by the genetic transformation with *LuTFL1* haplotypes. If *LuTFL1* has a significant impact on flowering time, growth determinacy and plant architecture, it will verify the model implemented in *PGROWTH* script

If, against the expectations, this approach fails to deliver positive results, the search for other candidate genes should continue. It was stated before that at the time of writing the list of genomic regions that might be under disruptive selection between fibre and oil flax has been identified (Soto-Cerda *et al.* 2013). This gives an excellent choice of genes that should be investigated in order to trace the evolution of flax fibre varieties. Phylogeographic approaches, followed by transgenic experiments would allow one to pin point the genes that played an important role in adaptation to northerly latitudes and in emergence of fibre varieties.

There is another, interesting conclusion drawn from this study that is beyond the scope of the posed hypothesis: it might be the case that long-standing adaptations that emerged during millions of years of wild ancestors evolution have advantage over simple, knock-out mutations that arose under high selection pressure during recent millennia of crop plants evolution. It might be the case that the former allow plants to be more flexible to environmental changes. Flax is a perfect model organism to test this hypothesis as its wild relatives distribution stretches through many climatic regions of Africa, Asia and Europe and in consequence these wild populations should be characterized by many long-standing adaptations to the local environments. To study this phenomenon one would have to investigate the local adaptations in both pale and cultivated flax, compare the mechanisms of this adaptation, its impact on fitness and evolvability.

BIBLIOGRAPHY

Achaz G (2013) http://wwwabi.snv.jussieu.fr/achaz/neutralitytest.html.

Alabadi D, Oyama T, Yanovsky MJ, *et al.* (2001) Reciprocal regulation between *TOC1* and *LHY*/*CCA1* within the *Arabidopsis* circadian clock. *Science* **293**, 880-883.

Allaby RG, Fuller Q, Kitchen J (2014) The limits of selection under plant domestication. http://arxiv.org/abs/1403.1244.

Allaby RG, Fuller DQ, Brown TA (2008) The genetic expectations of a protracted model for the origins of domesticated crops. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 13982-13986.

Allaby RG, Gutaker R, Clarke A, *et al.* (2014) Using archaeogenomic and computational approaches to unravel the history of local adaptation in crops. *Philosophical Transactions B*, (In press).

Allaby RG, Peterson GW, Merriwether DA, Fu YB (2005) Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the *SAD2* locus. *Theoretical and Applied Genetics* **112**, 58-65.

Alvarez J, Guli CL, Yu XH, Smyth DR (1992) *TERMINAL-FLOWER* - a gene affecting inflorescence development in *Arabidopsis thaliana*. *Plant Journal* **2**, 103-116.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-410.

Andrews S (2012) *FastQC A Quality Control tool for High Throughput Sequence Data.* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angelova S (2011): *Linum bienne* Mill., e-mail communication. Institute for Plant Genetic Resources, Bulgaria.

Azhaguvel P, Komatsuda T (2007) A phylogenetic analysis based on nucleotide sequence of a marker linked to the brittle rachis locus indicates a diphyletic origin of barley. *Annals of Botany* **100**, 1009-1015.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**.

Balfourier F, Imbert C, Charmet G (2000) Evidence for phylogeographic structure in *Lolium* species related to the spread of agriculture in Europe. A cpDNA study. *Theoretical and Applied Genetics* **101**, 131-138.

Barchi L, Lanteri S, Portis E*, et al.* (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *Bmc Genomics* **12**.

Barchi L, Lanteri S, Portis E*, et al.* (2012) A RAD tag derived marker based eggplant linkage map and the location of QTLs determining anthocyanin pigmentation. *PLoS One* **7**.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16**, 113-148.

Beerli P (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341-345.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4563-4568.

Blackman BK, Strasburg JL, Raduski AR, Michaels SD, Rieseberg LH (2010) The role of recently derived *FT* paralogs in sunflower domestication. *Current Biology* **20**, 629-635.

Bohlenius H, Huang T, Charbonnel-Campaa L*, et al.* (2006) *CO*/*FT* regulatory module controls timing of flowering and seasonal growth cessation in trees. *Science* **312**, 1040-1043.

Bollongino R, Nehlich O, Richards MP*, et al.* (2013) 2000 years of parallel societies in stone age central Europe. *Science* **342**, 479-481.

Boric D, Price TD (2013) Strontium isotopes document greater human mobility at the start of the Balkan Neolithic. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 3298-3303.

Bowman JL (1997) Evolutionary conservation of angiosperm flower development at the molecular and genetic levels. *Journal of Biosciences* **22**, 515-527.

Bradley D, Carpenter R, Copsey L*, et al.* (1996) Control of inflorescence architecture in *Antirrhinum*. *Nature* **379**, 791-797.

Bradley D, Ratcliffe O, Vincent C, Carpenter R, Coen E (1997) Inflorescence commitment and architecture in *Arabidopsis*. *Science* **275**, 80-83.

Bramanti B, Thomas MG, Haak W, *et al.* (2009) Genetic discontinuity between local hunter-gatherers and central europe's first farmers. *Science* **326**, 137-140.

Bremer B, Bremer K, Chase MW, *et al.* (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* **161**, 105-121.

Brown JCL, De Decker MM, Fieldes MA (2008) A comparative analysis of developmental profiles for DNA methylation in 5-azacytidine-induced early-flowering flax lines and their control. *Plant Science* **175**, 217-225.

Brown TA, Jones MK, Powell W, Allaby RG (2009) The complex origins of domesticated crops in the Fertile Crescent. *Trends in Ecology & Evolution* **24**, 103-109.

Bus A, Hecht J, Huettel B, Reinhardt R, Stich B (2012) High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *Bmc Genomics* **13**.

Cariou E, Brochard M, Raynal G (2003) La brulure du lin, due a Chalara elegans et plusieurs especes de Pythium. *Phytoma*, 6-9.

Carmel-Goren L, Liu YS, Lifschitz E, Zamir D (2003) The *SELF-PRUNING* gene family in tomato. *Plant Molecular Biology* **52**, 1215-1222.

Catanzariti A, Dodds PN, Ellis JG, Staskawicz BJ (2009) The interaction of avirulence and resistance gene products in flax rust disease. *Canadian Journal of Plant Pathology-Revue Canadienne De Phytopathologie* **31**, 502-502.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3-Genes Genomes Genetics* **1**, 171-182.

Chapman MA, Pashley CH, Wenzler J*, et al.* (2008) A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). *Plant Cell* **20**, 2931-2945.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**, 155-174.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A, *et al.* (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *Bmc Genomics* **12**.

Clarke JH, Dean C (1994) Mapping *FRI*, a locus controlling flowering time and vernalization response in *Arabidopsis thaliana*. *Molecular & General Genetics* **242**, 81-89.

Cockram J, Jones H, Leigh FJ, *et al.* (2007) Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. *Journal of Experimental Botany* **58**, 1231-1244.

Comadran J, Kilian B, Russell J, *et al.* (2012) Natural variation in a homolog of *Antirrhinum CENTRORADIALIS* contributed to spring growth habit and environmental adaptation in cultivated barley. *Nature Genetics* **44**, 1388-1392.

Concolly J, Colledge S, Shennan S (2008) Founder effect, drift, and adaptive change in domestic crop use in early Neolithic Europe. *Journal of Archaeological Science* **35**, 2797-2804.

Coward F, Shennan S, Colledge S, Conolly J, Collard M (2008) The spread of Neolithic plant economies from the Near East to northwest Europe: a phylogenetic analysis. *Journal of Archaeological Science* **35**, 42-56.

Darapuneni MK, Morgan GD, Ibrahim AMH, Duncan RW (2014) Effect of vernalization and photoperiod on flax flowering time. *Euphytica* **195**, 279-285.

Davey JW, Blaxter ML (2011) RADSeq: next-generation population genetics (vol 9, pg 416, 2010). *Briefings in Functional Genomics* **10**, 108-108.

Davis P (1988) Linaceae. In: *Flora of Turkey*. Edinburgh University Press, Edinburgh.

De Decker MM (2007) *Studies of DNA Methylation and Flowering Time Genes in Early Flowering Flax (Linum usitatissimum L.) Lines Induced by 5-Azacytidine*, University of Waterloo.

de Wit PJGM, Mehrabi R, van den Burg HA, Stergiopoulos I (2009) Fungal effector proteins: past, present and future. *Molecular Plant Pathology* **10**, 735-747.

Debieu M, Tang C, Stich B*, et al.* (2013) Co-variation between seed dormancy, growth rate and flowering time changes with latitude in *Arabidopsis thaliana*. *PLoS One* **8**.

Diaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS One* **7**.

Diederichsen A (2009) Taxonomical and morphological assessment of infra-specific diversity in cultivated flax (*Linum usitatissimum* L.). *Journal of Agriculture and Rural Development in the Tropics and Subtropics*, 33-51.

Diederichsen A, Fu YB (2006) Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*). *Genetic Resources and Crop Evolution* **53**, 77-90.

Diederichsen A, Hammer K (1995) Variation of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*) and its wild progenitor pale flax (subsp. *angustifolium* (Huds) Thell). *Genetic Resources and Crop Evolution* **42**, 263-272.

Diederichsen A, Raney JP (2006) Seed colour, seed weight and seed oil content in *Linum usitatissimum* accessions held by Plant Gene Resources of Canada. *Plant Breeding* **125**, 372-377.

Diederichsen A, Ulrich A (2009) Variability in stem fibre content and its association with other characteristics in 1177 flax (*Linum usitatissimum* L.) genebank accessions. *Industrial Crops and Products* **30**, 33-39.

Domantovich A, Koshkin V, Brutch N, Matvienko I (2012) Investigation of photoperiod sensitivity of *Linum usitatissimum* L. Lines and effect of short-day conditions on their economically valuable traits. *Russian Agricultural Sciences* **38**, 173-177.

Dribnenki P (2010) Annual report: Northern adapted flax development project.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**, 359-361.

Ehrenreich IM, Hanzawa Y, Chou L, *et al.* (2009) Candidate gene association mapping of *Arabidopsis* flowering time. *Genetics* **183**, 325-335.

Elitzur T, Nahum H, Borovsky Y, et al. (2009) Co-ordinated regulation of flowering time, plant architecture and growth by *FASCICULATE*: the pepper ortholog of *SELF PRUNING*. *Journal of Experimental Botany* **60,** 869-880.

Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6.**

Etter PD, Bassham S, Hohenlohe Pl, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (ed. Rockman MV), pp. 157-178. Humana Press, Orgogozo, Virginie

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Everaert I, De Riek J, De Loose M, Van Waes J, Van Bockstaele E (2001) Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. *Plant Varieties and Seeds* **14**, 69-87.

Eyre-Walker A, Gaut RL, Hilton H, Feldman DL, Gaut BS (1998) Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 4441-4446.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567-1587.

Fieldes MA, Amyot LM (1999) Epigenetic control of early flowering in flax lines induced by 5-azacytidine applied to germinating seed. *Journal of Heredity* **90**, 199-206.

Fieldes MA, Harvey CG (2004) Differences in developmental programming and node number at flowering in the 5-azacytidine-induced early flowering flax lines and their controls. *International Journal of Plant Sciences* **165**, 695-706.

Fjellheim S, Rognli OA, Fosnes K, Brochmann C (2006) Phylogeographical history of the widespread meadow fescue (*Festuca pratensis* Huds.) inferred from chloroplast DNA sequences. *Journal of Biogeography* **33**, 1470-1478.

Flavell AJ, Knox MR, Pearce SR and Ellis THN (1999) Retrotransposon-based insertion polymorphisms (RBIP) for high-throughput marker analysis. *Plant Journal* **16**, 643-650.

Flowers JM, Hanzawa Y, Hall MC, Moore RC, Purugganan MD (2009) Population genomics of the *Arabidopsis thaliana* flowering time gene network. *Molecular Biology and Evolution* **26**, 2475-2486.

Fu Y-B (2011) Genetic evidence for early flax domestication with capsular dehiscence. *Genetic Resources and Crop Evolution* **58**, 1119-1128.

Fu Y-B (2012) Population-based resequencing revealed an ancestral winter group of cultivated flax: implication for flax domestication processes. *Ecology and Evolution* **2**, 622-635.

Fu Y-B, Allaby RG (2010) Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. *Genetic Resources and Crop Evolution* **57**, 667-677.

Fu Y-B, Diederichsen A, Allaby RG (2012) Locus-specific view of flax domestication history. *Ecology and Evolution* **2**, 139-152.

Fu Y-B, Peterson GW (2010) Characterization of expressed sequence tag-derived simple sequence repeat markers for 17 *Linum* species. *Botany-Botanique* **88**, 537-543.

Fu Y-B, Peterson GW (2012) Developing genomic resources in two *Linum* species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources* **12**, 492-500.

Fu YX, Li WH (1993) Statistical test of neutrality of mutations. *Genetics* **133**, 693-709.

Fuller DQ, Allaby RG (2009) Seed Dispersal and Crop Domestication: shattering, germination and seasonality in evolution under cultivation. In: *Fruit Development and Seed Dispersal.* (ed. Ostergaard L), pp. 238 - 295. Blackwell, Oxford.

Fuller DQ, Willcox G, Allaby RG (2012) Early agricultural pathways: moving outside the 'core area' hypothesis in Southwest Asia. *Journal of Experimental Botany* **63**, 617-633.

Gao H, Williamson S, Bustamante CD (2007) A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics* **176**, 1635-1651.

Gascuel O (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**, 685-695.

Gelman A, Rubin D (1992) Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-472.

Gepts P (2004) Crop domestication as a long-term selection experiment. *Plant Breeding Review* **24**, 1-44.

Gill KS, Yermanos DM (1967) Cytogenetic studies on the genus *Linum* L. hybrids among taxa with 15 as the Haploid Chromosome Number. *Crop Science* **7**, 623–627.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**, 184-186.

Gurbuz B (1999) Determination of cross-pollination in flax (*Linum usitatissimum*) using different experimental designs. *Journal of Agricultural Science* **133**, 31-35.

Haak W, Balanovsky O, Sanchez JJ, *et al.* (2010) Ancient DNA from european early Neolithic farmers reveals their Near Eastern affinities. *PLoS Biology* **8**.

Hall D, Ma X-F, Ingvarsson PK (2011) Adaptive evolution of the *Populus tremula* photoperiod pathway. *Molecular Ecology* **20**, 1463-1474.

Hammer K (1984) Das Domestikationssyndrom. *Kulturpflanze* **32**, 11-34.

Hammer K (1986) Linaceae. In: *Verzeichnis landwirtshaftlicher und gartenrischer Kulturpflanzen.* (ed. Schultze-Motel J), pp. 710-713. Akademie-Verlag, Berlin.

Hayama R, Coupland G (2004) The molecular basis of diversity in the photoperiodic flowering responses of Arabidopsis and rice. *Plant Physiology* **135**, 677-684.

Helbaek H (1959) Domestication of the plants in the old world. *Science* **130**, 365-372.

Herbig C, Maier U (2011) Flax for oil or fibre? Morphometric analysis of flax seeds and new aspects of flax cultivation in Late Neolithic wetland settlements in southwest Germany. *Vegetation History and Archaeobotany* **20**, 527-533.

Hervella M, Izagirre N, Alonso S, *et al.* (2012) Ancient DNA from hunter-gatherer and farmer groups from northern Spain supports a random dispersion model for the neolithic expansion into Europe. *PLoS One* **7**.

Hillman GC (1975) The plant remains from Tell Abu Hureyra: a preliminary report. *Proceedings of the Prehistoric Society* **41**, 70-73.

Hillman GC, Davies MS (1990) Domestication rates in wild-type wheats and barley under primitive cultivation. *Biological Journal of the Linnean Society* **39**, 39-78.

House AM (2010) *Molecular Studies of 5-Azacytidine-Induced Early- Flowering Lines of Flax*, Wilfrid Laurier University.

Holmgren PK, Holmgren NH (1998, continuously updated). *Index herbariorum: a global directory of public herbaria and associated staff*. New York: New York Botanical Garden. Retrieved 2014-09-25.

Hsu C-Y, Adams JP, Kim H, *et al.* (2011) *FLOWERING LOCUS T* duplication coordinates reproductive and vegetative growth in perennial poplar. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10756-10761.

Huang N-C, Jane W-N, Chen J, Yu T-S (2012) Arabidopsis thaliana CENTRORADIALIS homologue (ATC) acts systemically to inhibit floral initiation in Arabidopsis. *Plant Journal* **72**, 175-184.

Hubner S, Guenther T, Flavell A, *et al.* (2012) Islands and streams: clusters and gene flow in wild barley populations from the Levant. *Molecular Ecology* **21**, 1115-1129.

Huis R, Hawkins S, Neutelings G (2010) Selection of reference genes for quantitative gene expression normalization in flax (Linum usitatissimum L.). *Bmc Plant Biology* **10**.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254-267.

Igasaki T, Watanabe Y, Nishiguchi M, Kotoda N (2008) The *FLOWERING LOCUS T/TERMINAL FLOWER 1* family in Lombardy poplar. *Plant and Cell Physiology* **49**, 291-300.

Iqbal M, Navabi A, Salmon DF, *et al.* (2007) Genetic analysis of flowering and maturity time in high latitude spring wheat - Genetic analysis of earliness in spring wheat. *Euphytica* **154**, 207-218.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070-3071.

Jones G, Buck C, Charles M, Higham T, Colledge S (2007) Out of Asia - the spread of agriculture. In: *Planet Earth Online*, pp. 22-23

Jones G, Jones H, Charles MP, *et al.* (2012) Phylogeographic analysis of barley DNA as evidence for the spread of Neolithic agriculture through Europe. *Journal of Archaeological Science* **39**, 3230–3238.

Jones H, Civan P, Cockram J, *et al.* (2011) Evolutionary history of barley cultivation in Europe revealed by genetic analysis of extant landraces. *Bmc Evolutionary Biology* **11**.

Jones H, Leigh FJ, Mackay I, *et al.* (2008) Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the fertile crescent. *Molecular Biology and Evolution* **25**, 2211-2219.

Kalendar R, Grob T, Regina M, Suomeni A, Schulman A. (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. *Theoretical and Applied Genetics* **98**, 704–711.

Kalia S, Kaith BS, Kaur I (2009) Pretreatments of natural fibers and their application as reinforcing material in polymer composites - a review. *Polymer Engineering and Science* **49**, 1253-1272.

Kardailsky I, Shukla VK, Ahn JH, *et al.* (1999) Activation tagging of the floral inducer FT. *Science* **286**, 1962-1965.

Karg S (2011) New research on the cultural history of the useful plant *Linum usitatissimum* L. (flax), a resource for food and textiles for 8,000 years. *Vegetation History and Archaeobotany* **20**, 507-508.

Keller SR, Levsen N, Ingvarsson PK, Olson MS, Tiffin P (2011a) Local selection across a latitudinal gradient shapes nucleotide diversity in balsam poplar, *Populus balsamifera* L. *Genetics* **188**, 941-952.

Keller SR, Soolanayakanahally RY, Guy RD, *et al.* (2011b) Climate-driven local adaptation of ecophysiology and phenology in balsam poplar, *Populus balsamifera* L. (Salcaceae). *American Journal of Botany* **98**, 99-108.

Kilian B, Ozkan H, Walther A, *et al.* (2007) Molecular diversity at 18 loci in 321 wild and 92 domesticate lines reveal no reduction of nucleotide diversity during *Triticum monococcum* (einkorn) domestication: Implications for the origin of agriculture. *Molecular Biology and Evolution* **24**, 2657-2668.

Kirk JTO (1994) *Light and photosynthesis in aquatic ecosystems*, 2 edn. Cambridge University Press, Cambridge.

Kobayashi Y, Kaya H, Goto K, Iwabuchi M, Araki T (1999) A pair of related genes with antagonistic roles in mediating flowering signals. *Science* **286**, 1960-1962.

Koornneef M, Blankestijndevries H, Hanhart C, Soppe W, Peeters T (1994) The phenotype of some late-flowering mutants is enhanced by a locus on chromosome 5 that is not effective in the Landsberg *erecta* wild-type. *Plant Journal* **6**, 911-919.

Kreuz A, Marinova E, Schafer E, Wiethold J (2005) A comparison of early Neolithic crop and weed assemblages from the Linearbandkeramik and the Bulgarian Neolithic cultures: differences and similarities. *Vegetation History and Archaeobotany* **14**, 237-258.

Krieger U, Lippman ZB, Zamir D (2010) The flowering gene *SINGLE FLOWER TRUSS* drives heterosis for yield in tomato. *Nature Genetics* **42**, 459-463.

Kulpa W, Danert S (1962) Zur systematik von *Linum usitatissimum* L. *Kulturpflanzen* **3**, 341-388.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359.

Larkin MA, Blackshields G, Brown NP, *et al.* (2007) Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947-2948.

Laurie RE, Diwadkar P, Jaudal M, *et al.* (2011) The *Medicago FLOWERING LOCUS T* homolog, *MtFTa1*, is a key regulator of flowering time. *Plant Physiology* **156**, 2207-2224.

Leff B, Ramankutty N, Foley JA (2004) Geographic distribution of major crops across the world. *Global Biogeochemical Cycles* **18**, 1-27.

Leseberg CH, Li A, Kang H, Duvall M, Mao L (2006) Genome-wide analysis of the MADS-box gene family in Populus trichocarpa. *Gene* **378**, 84-94.

Leuzinger U, Rast-Eicher A (2011) Flax processing in the Neolithic and Bronze Age pile-dwelling settlements of eastern Switzerland. *Vegetation History and Archaeobotany* **20**, 535-542.

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; Heterotic models. *Genetics* **49**, 49-67.

Li B, Suzuki JI, Hara T (1998) Latitudinal variation in plant size and relative growth rate in *Arabidopsis thaliana*. *Oecologia* **115**, 293-301.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Lifschitz E, Eshed Y (2006) Universal florigenic signals triggered by *FT* homologues regulate growth and flowering cycles in perennial day-neutral tomato. *Journal of Experimental Botany* **57**, 3405-3414.

Lifschitz E, Eviatar T, Rozman A, *et al.* (2006) The tomato *FT* ortholog triggers systemic signals that regulate growth and flowering and substitute for diverse environmental stimuli. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6398-6403.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298-299.

Lister DL, Bower MA, Jones MK (2010) Herbarium specimens expand the geographical and temporal range of germplasm data in phylogeographic studies. *Taxon* **59**, 1321-1323.

Lister DL, Thaw S, Bower MA, *et al.* (2009) Latitudinal variation in a photoperiod response gene in European barley: insight into the dynamics of agricultural spread from 'historic' specimens. *Journal of Archaeological Science* **36**, 1092-1098.

Maier U, Schlichtherle H (2011) Flax cultivation and textile production in Neolithic wetland settlements on Lake Constance and in Upper Swabia (south-west Germany). *Vegetation History and Archaeobotany* **20**, 567-578.

Mandel MA, Gustafsonbrown C, Savidge B, Yanofsky MF (1992) Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. *Nature* **360**, 273-277.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

Matsoukas IG, Massiah AJ, Thomas B (2012) Florigenic and antflorigenic signaling in plants. *Plant and Cell Physiology* **53**, 1827-1842.

McDill J, Repplinger M, Simpson BB, Kadereit JW (2009) The phylogeny of *Linum* and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. *Systematic Botany* **34**, 386-405.

McDill JR, Simpson BB (2011) Molecular phylogenetics of Linaceae with complete generic sampling and data from two plastid genes. *Botanical Journal of the Linnean Society* **165**, 64-83.

McGarry RC, Ayre BG (2012a) Geminivirus-mediated delivery of florigen promotes determinate growth in aerial organs and uncouples flowering from photoperiod in cotton. *PLoS One* **7**.

McGarry RC, Ayre BG (2012b) Manipulating plant architecture with members of the CETS gene family. *Plant Science* **188**, 71-81.

Mimida N, Goto K, Kobayashi Y*, et al.* (2001) Functional divergence of the *TFL1*-like gene family in *Arabidopsis* revealed by characterization of a novel homologue. *Genes to Cells* **6**, 327-336.

Mizoguchi T, Wright L, Fujiwara S*, et al.* (2005) Distinct roles of *GIGANTEA* in promoting flowering and regulating circadian rhythms in *Arabidopsis*. *Plant Cell* **17**, 2255-2270.

Munguia-Rosas MA, Ollerton J, Parra-Tabla V, Arturo De-Nova J (2011) Meta-analysis of phenotypic selection on flowering phenology suggests that early flowering plants are favoured. *Ecology Letters* **14**, 511-521.

Muravenko OV, Lemesh VA, Samatadze TE*, et al.* (2003) Genome comparisons with chromosomal and molecular markers for three closely related flax species and their hybrids. *Russian Journal of Genetics* **39**, 510-518.

National Geospatial-Intelligence Agency (2012) https://http://www.cia.gov/library/publications/the-world-factbook/fields/2011.html.

Nei M (1977) F-statistics and analysis o gene diversity in subdivided populations. *Annals of Human Genetics* **41**, 225-233.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 5269-5273.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics* **39**, 197-218.

Nikolić T ed. (2010): *Linum bienne* Mill.. Flora Croatica Database (URL http://hirc.botanic.hr/fcd/). Faculty of Science, University of Zagreb, Croatia.

Ozkan H, Brandolini A, Pozzi C*, et al.* (2005) A reconsideration of the domestication geography of tetraploid wheats. *Theoretical and Applied Genetics* **110**, 1052-1060.

Ozkan H, Willcox G, Graner A, Salamini F, Kilian B (2011) Geographic distribution and domestication of wild emmer wheat (*Triticum dicoccoides*). *Genetic Resources and Crop Evolution* **58**, 11-53.

Palmer SA, Moore JD, Clapham AJ, Rose P, Allaby RG (2009) Archaeogenetic evidence of ancient Nubian barley evolution from six to two-row indicates local adaptation. *PLoS One* **4**.

Paradis E (2010) Pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* **26**, 419-420.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290.

Park DH, Somers DE, Kim YS*, et al.* (1999) Control of circadian rhythms and photoperiodic flowering by the *Arabidopsis GIGANTEA* gene. *Science* **285**, 1579-1582.

Pfender WF, Saha MC, Johnson EA, Slabaugh MB (2011) Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. *Theoretical and Applied Genetics* **122**, 1467-1480.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**.

Pin PA, Benlloch R, Bonnet D*, et al.* (2010) An antagonistic pair of *FT* homologs mediates the control of flowering time in sugar beet. *Science* **330**, 1397-1400.

Planchon JE (1848) Sur la famille des Linees. *The London Journal of Botany* **7**, 165-186, 473-501, 507-528.

Price TD (2000) *Europe's first farmers* Cambridge University Press, Cambridge.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* **69**, 1-14.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Prusinkiewicz P, Erasmus Y, Lane B, Harder LD, Coen E (2007) Evolution and development of inflorescence architectures. *Science* **316**, 1452-1456.

Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. *Nature* **457**, 843-848.

Purugganan MD, Fuller DQ (2011) Archaeological data reveal slow rates of evolution during plant domestication. *Evolution* **65**, 171-183.

Putterill J (2001) Flowering in time: genes controlling photoperiodic flowering in *Arabidopsis*. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **356**, 1761-1767.

Putterill J, Robson F, Lee K, Simon R, Coupland G (1995) The *CONSTANS* gene of *Arabidopsis* promotes flowering and encodes a protein showing similarities to zinc-finger transcription factors. *Cell* **80**, 847-857.

Queller DC, Strassman JE, Hughes CR (1993) Microsatellites and kinship. *Trends in Ecology and Evolution* **8,** 285-288.

Ramos-Onsins SE, Rozas J (2002) Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution* **19**, 2092-2100.

Ronquist F, Teslenko M, van der Mark P*, et al.* (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**, 539-542.

Roux F, Touzet P, Cuguen J, Le Corre V (2006) How to be early flowering: an evolutionary perspective. *Trends in Plant Science* **11**, 375-381.

Rowland G, Cloutier S (2012) Linum.ca database, http://anther.ccis.ualberta.ca/ - a.

Samach A, Onouchi H, Gold SE*, et al.* (2000) Distinct roles of *CONSTANS* target genes in reproductive development of *Arabidopsis*. *Science* **288**, 1613-1616.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441–448.

Scaglione D, Acquadro A, Portis E*, et al.* (2012) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *Bmc Genomics* **13**.

Schaffer R, Ramsay N, Samach A*, et al.* (1998) The late elongated hypocotyl mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* **93**, 1219-1229.

Schoentgen F, Saccoccio F, Jolles J, Bernier I, Jolles P (1987) Complete amino-acid sequence of a basic 21-kDa protein from bovine brain cytosol. *European Journal of Biochemistry* **166**, 333-338.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Shalit A, Rozman A, Goldshmidt A*, et al.* (2009) The flowering hormone florigen functions as a general systemic regulator of growth and termination. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8392-8397.

Shannon S, Meekswagner DR (1991) A mutation in the *Arabidopsis TFL1* gene affects inflorescence mersitem development. *Plant Cell* **3**, 877-892.

Sheldon CC, Burn JE, Perez PP*, et al.* (1999) The *FLF* MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell* **11**, 445-458.

Sheldon CC, Rouse DT, Finnegan EJ, Peacock WJ, Dennis ES (2000) The molecular basis of vernalization: The central role of *FLOWERING LOCUS C* (*FLC*). *Proceedings of the National Academy of Sciences of the United States of America* **97**, 3753-3758.

Simon R, Igeno MI, Coupland G (1996) Activation of floral meristem identity genes in *Arabidopsis*. *Nature* **384**, 59-62.

Skoglund P, Malmstrom H, Raghavan M*, et al.* (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466-469.

Skorkowska-Telichowska K, Zuk M, Kulma A*, et al.* (2010) New dressing materials derived from transgenic flax products to treat long-standing venous ulcers-a pilot study. *Wound Repair and Regeneration* **18**, 168-179.

Smartt J (1990) *Grain Legumes: Evolution and Genetic Resources* Cambridge University Press, Cambridge.

Smith JM, Haigh J (2007) The hitch-hiking effect of a favourable gene. *Genetics Research* **89**, 391-403.

Smykal P, Bacova-Kerteszova N, Kalendar R*, et al.* (2011) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. *Theoretical and Applied Genetics* **122**, 1385-1397.

Solar PM (2012) The triumph of cotton in Europe. In: *Modern and Comparative Seminar*. London School of Economics and Political Science.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* **7**, 1-12.

Soto-Cerda BJ, Cloutier S (2013) Outlier loci and selection signature of simple sequence repeats (SSRs) in flax (*Linum usitatissimum* L.). *Plant Molecular Biology Reporter* **31,** 978-990.

Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S (2013) Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. *Bmc Plant Biology* **13**.

Soto-Cerda BJ, Maureira-Butler I, Munoz G, Rupayan A, Cloutier S (2012) SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. *Molecular Breeding* **30**, 875-888.

Stoelting KN, Nipper R, Lindtke D*, et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology* **22**, 842-855.

Strid A ed (2011): *Linum bienne* Mill., Flora Hellenica Database. Univeristy of Copenhagen, Denmark.

Suarez-Lopez P, Wheatley K, Robson F*, et al.* (2001) *CONSTANS* mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature* **410**, 1116-1120.

Tajima F (1989) Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595.

Takahashi R (1972) Non brittle rachis 1 and non brittle rachis 2. *Barley Genetic Newsletter* **2**, 181-182.

Takenaka S, Kawahara T (2012) Evolution and dispersal of emmer wheat (*Triticum* sp.) from novel haplotypes of *PPD-1* (photoperiod response) genes and their surrounding DNA sequences. *Theoretical and Applied Genetics* **125**, 999-1014.

Thomas B, Carre I, Jackson S (2006) Photoperiodism and flowering. In: *The Molecular Biology and Biotechnology of Flowering* (ed. Jordan BR). CAB International, Wallingford.

Tian Z, Wang X, Lee R*, et al.* (2010) Artificial selection for determinate growth habit in soybean. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8563-8568.

Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604.

Tolar T, Jacomet S, Veluscek A, Cufar K (2011) Plant economy at a late Neolithic lake dwelling site in Slovenia at the time of the Alpine Iceman. *Vegetation History and Archaeobotany* **20**, 207-222.

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635-641.

Tsaftaris A, Pasentsis K, Kalivas A*, et al.* (2012) Isolation of a *CENTRORADIALIS*/*TERMINAL FLOWER1* homolog in saffron (*Crocus sativus* L.): characterization and expression analysis. *Molecular Biology Reports* **39**, 7899-7910.

Uysal H, Fu YB, Kurt O*, et al.* (2010) Genetic diversity of cultivated flax (Linum usitatissimum L.) and its wild progenitor pale flax (Linum bienne Mill.) as revealed by ISSR markers. *Genetic Resources and Crop Evolution* **57**, 1109-1119.

Uysal H, Kurt O, Fu Y-B, Diederichsen A, Kusters P (2012) Variation in phenotypic characters of pale flax (*Linum bienne* Mill.) from Turkey. *Genetic Resources and Crop Evolution* **59**, 19-30.

Valamoti SM (2011) Flax in Neolithic and Bronze Age Greece: archaeobotanical evidence. *Vegetation History and Archaeobotany* **20**, 549-560.

Van Zeist W, Bakker-Heeres JAH (1975) Evidence for linseed cultivation before 6000 BC. *Journal of Archaeological Sciences* **2**, 215-219.

Vavilov NI (1951) *The origin, immunity and breeding of cultivated plants* Waltham, Mass. and Wm. Dawson and Sons, London.

Vereshch AG, Novitska GV (1965) Triglyceride composition of linseed oil. *Journal of the American Oil Chemists Society* **42**, 970-974.

Vos P, Hogers R, Bleeker M, *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**, 4407–4414.

Vromans J (2006) *Molecular genetic studies in flax (Linum usitatissimum L.)*, Wageningen University.

Vul'f EV, Elladi EV (1940) Familie Linaceae (D.C.) Dumort-Leinewachse (Russ.). In: *Kul'turnaja Flora SSSR (Faserpflanzen Teil 1)* (ed. Vul'f EV), pp. 97-207, Moscow-St. Petersburg.

Walsh B (2008) Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica* **161**, 1-17.

Wang L-N, Liu Y-F, Zhang Y-M, Fang R-X, Liu Q-L (2012a) The expression level of *Rosa Terminal Flower 1* (*RTFL1*) is related with recurrent flowering in roses. *Molecular Biology Reports* **39**, 3737-3746.

Wang N, Fang L, Xin H, Wang L, Li S (2012b) Construction of a high-density genetic map for grape using next generation restriction-site associated DNA sequencing. *Bmc Plant Biology* **12**.

Wang N, Thomson M, Bodles WJA, *et al.* (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology* **22**, 3098-3111.

Wang ZY, Tobin EM (1998) Constitutive expression of the *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) gene disrupts circadian rhythms and suppresses its own expression. *Cell* **93**, 1207-1217.

Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM (1992) *LEAFY* controls floral meristem identity in *Arabidopsis*. *Cell* **69**, 843-859.

Weller JL, Liew LC, Hecht VFG, *et al.* (2012) A conserved molecular basis for photoperiod adaptation in two temperate legumes. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 21158-21163.

Winkler H (1931) Linaceae. In: *Die naturalichen Pflanzenfamilien* (ed. Engler HGA, Prantl, K. A. E.), pp. 67-112. Engelman, Leipzig.

Wright S (1950) Genetical structure of populations. *Nature* **166**, 247-249.

Yamaguchi A, Kobayashi Y, Goto K, Abe M, Araki T (2005) *TWIN SISTER OF FT* (*TSF*) acts as a floral pathway integrator redundantly with *FT*. *Plant and Cell Physiology* **46**, 1175-1189.

Yamasaki M, Tenaillon MI, Bi IV*, et al.* (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859-2872.

Yang H, Tao Y, Zheng Z*, et al.* (2012) Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in *Lupinus angustifolius* L. *Bmc Genomics* **13**.

Yermanos DM, Gill KS (1969) Cytology of autotetraploids of *Linum usitatissimum* L. and *Linum angustifolium* Huds. and their amphidiploid hybrids. *Crop Science* **9**, 249-250.

Yoo SJ, Chung KS, Jung SH*, et al.* (2010) *BROTHER OF FT AND TFL1* (*BFT*) has *TFL1*-like activity and functions redundantly with *TFL1* in inflorescence meristem development in *Arabidopsis*. *Plant Journal* **63**, 241-253.

Yoo SY, Kardailsky I, Lee JS, Weigel D, Ahn JH (2004) Acceleration of flowering by overexpression of *MFT* (*MOTHER OF FT AND TFL1*). *Molecules and Cells* **17**, 95-101.

Yuceer C, Harkess RL, Land SB, Luthe DS (2002) Structure and developmental regulation of *CONSTANS-LIKE* genes isolated from *Populus deltoides*. *Plant Science* **163**, 615-625.

Zhan ZP, Huang FR, Luo J*, et al.* (2009) Duration of feeding linseed diet influences expression of inflammation-related genes and growth performance of growing-finishing barrows. *Journal of Animal Science* **87**, 603-611.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Molecular Biology and Evolution* **24**, 875-888.

Zohary D (1999) Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. *Genetic Resources and Crop Evolution* **46**, 133-142.

Zohary D, Hopf M (2000) *Domestication of plants in the Old World: The origin and spread of cultivated plants in West Asia, Europe, and the Nile Valley.*, 3rd edn. Oxford University Press, Oxford.

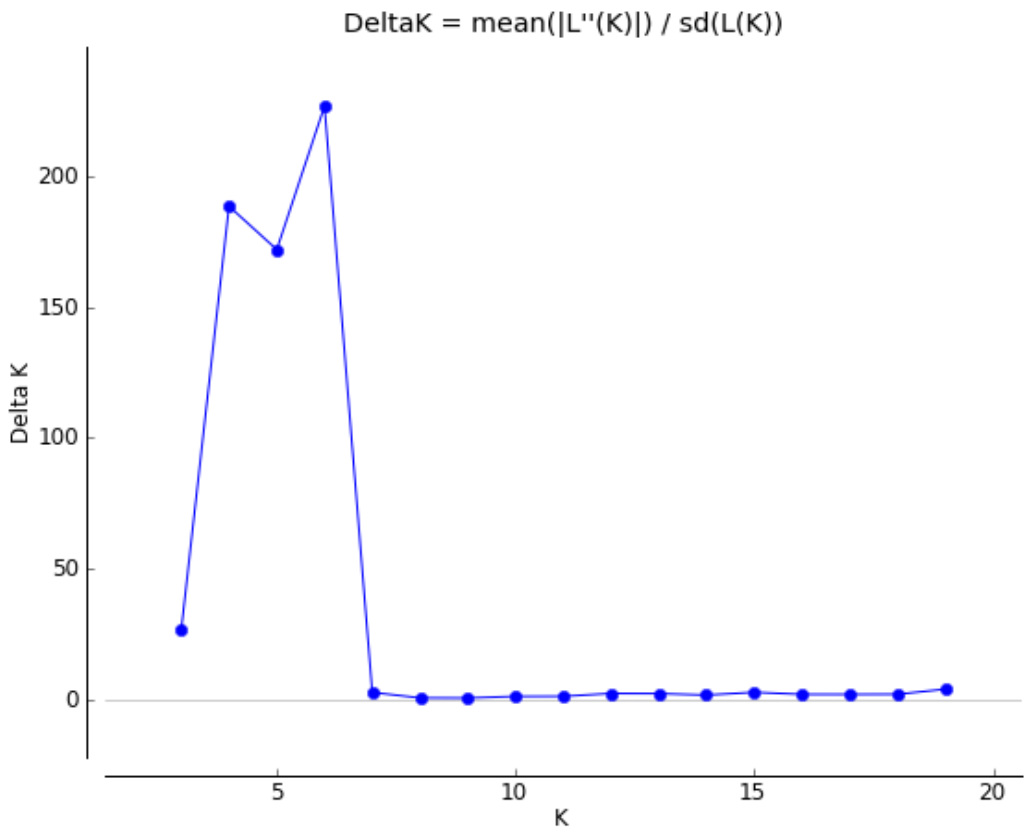
Supplementary information

Supplement 1: List of accessions used in Chapter 4 with latitude and *LuTFL* haplotypes.

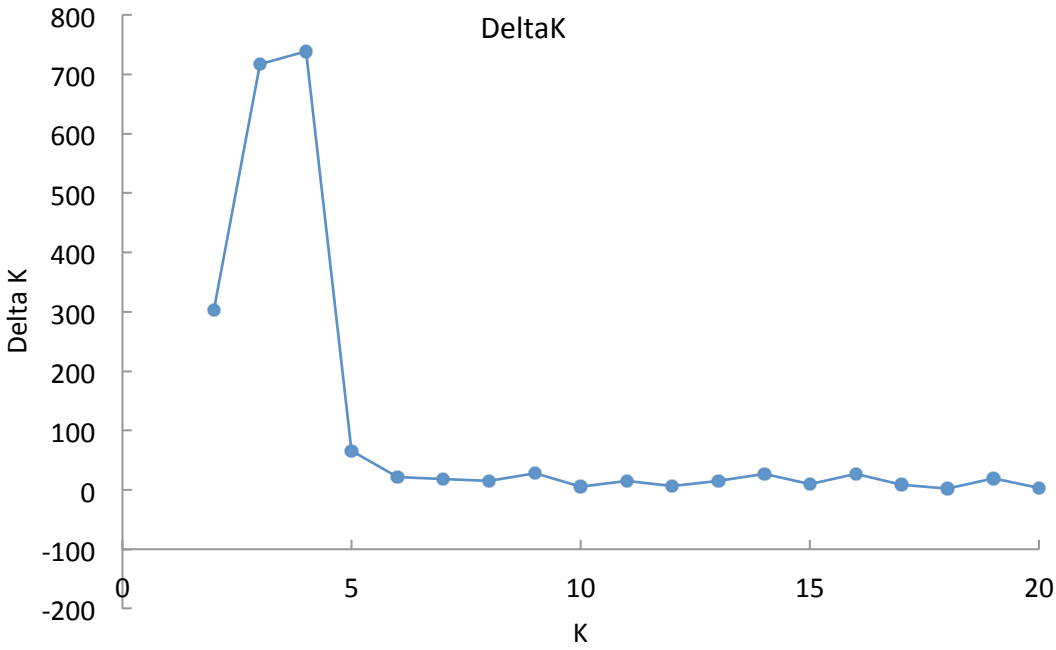
Accession No.	Latitude	*LuTFL1* haplotype	*LuTFL2* haplotype
M001	60	I	VII
M002	39	?	?
M003	39,5	I	VII
M004	40	I	VII
M005	31,5	I	VII
M006	31,5	XII	VII
M007	32	I	VII
M008	32	XV	VII
M009	32	XIV	?
M010	39	I	?
M011	39	I	VI
M012	39	I	VII
M013	47	I	VII
M014	47	III	XI
M015	47	I	VII
M016	51	III	VII
M017	51	III	III
M018	52	III	XI
M019	52	III	VII
M020	52	I	VII
M021	49,5	?	XI
M022	49,5	XII	?
M023	49,5	III	VII
M024	60	III	XI
M025	60	I	XI
M026	60	III	XI
M027	49	III	XI
M028	49	XII	XI
M029	48	III	?
M030	46	III	XI
M031	46	III	XI
M032	46	XII	XI
M033	40	XII	IX
M034	32	XII	VII
M035	32	III	VIII
M036	34	?	XI
M037	34	I	X
M038	52,5	III	XI
M039	52,5	III	XI
M040	52,5	III	XI
M041	31,5	III	XI
M042	32	I	III
M043	39	I	?
M044	39	I	VII
M045	60	III	XI
M046	60	III	XI
M047	35,5	III	XI
M048	27	XII	VII
M049	34	I	XI
M050	32	I	X
M051	32	I	?
M052	39	III	X
M053	39	I	III
M054	43	XII	VII
M055	43	XII	VII
M056	46	?	VII
M057	46	?	VII
M058	47	XII	VII
M059	47	I	IV
M060	51	III	XI
M061	51	III	XI
M062	52	III	XI
M063	49,5	I	VII
M064	49,5	III	XI
M065	46	VIII	XI
M066	46	III	III
M067	54	I	VII
M068	40	?	VII
M069	27	III	XI
M070	27	III	XI

M071	32	III	?
M072	60	XIII	VI
M073	48	VI	XI
M074	49	XII	XI
M075	49	XII	XI
M076	48	XI	VII
H002	52,167	III	XI
H009	52	III	XI
H012	52,2	XVI	XI
H013	52	III	XI
H031	54	?	?
H042	42	I	VI
H047	49,267	I	XI
H048	47,133	IX	V
H050	37	I	III
H051	35	I	VI
H083	55,167	III	?
H084	53,017	III	XI
H085	49,717	III	XI
H086	43	I	XI
H087	48,667	III	XI
H088	40,75	III	VII
H089	40,8	III	?
H090	57,35	III	XI
H091	57,017	III	XI
H092	55,183	III	XI
H093	38,5	III	VII
H094	38,833	III	XI
H095	38,833	?	?
H096	35,25	III	XI
H098	34	?	VI
H100	39,9	?	?
H101	39,217	I	?
H102	57,15	?	XI
H103	60	III	XI
H104	52,5	III	XI
H105	52,5	III	XI
H106	39	III	?
H107	32,05	I	III
H108	57,15	III	VII
H109	51	III	XI
H110	46	III	XI
H111	51,683	III	XI
H112	45,333	III	VII
W042	37,567	I	I
W043	41,183	III	I
W044	39	II	I
W045	39	II	I
W046	46	VIII	III
W047	46	VIII	III
W048	41,35	I	XI
W049	41,383	I	I
W050	37,817	I	III
W051	37,033	I	I
W052	41,583	X	III
W053	41,217	III	I
W054	41,833	III	?
W055	40,433	III	I
W056	40,817	IX	I
W057	40,75	I	XII
W065	43,9	VIII	III
W066	43,467	III	III
W067	43,017	III	I
W068	42,233	IV	I
W069	41,167	III	I
W070	39,9	VI	II
W072	38,983	V	II
W074	39,05	?	I
W076	39,783	II	I
W077	39,967	II	I
W081	40,35	III	I
W082	40,583	III	I
W085	41,583	VII	III
W086	41,556		
W094	45,13	VIII	III
W095	45,367	VIII	III
W096	45,417	VIII	III

Supplement 2: Graph summarizing likelihoods for different K populations in *STRUCTURE* analyses.



Supplement 3: Graph summarizing likelihoods for different K populations in *INSTRUCT* analyses.



List of electronic supplementary data

Data files are stored in the attached CD disk.

Electronic supplement 1: Multiple alignments for flowering time loci with designed degenerate primers.

Electronic supplement 2: Chromatograms for sequences of putative flowering time gene homologs in flax.

Electronic supplement 3: Script written in *R* conducting the permutational autocorrelation test.

Electronic supplement 4: Script written in *PERL* simulating the selection pressure on cultivated flax during the adaptation to the northerly latitudes.

Electronic supplement 5: Chromatograms for sequences of *LuTFL1* loci in flax.

Electronic supplement 6: Chromatograms for sequences of *LuTFL2* loci in flax.

Electronic supplement 7: Multiple alignment for *LuTFL1* sequences in flax.

Electronic supplement 8: Multiple alignment for *LuTFL2* sequences in flax.

Electronic supplement 9: Raw linkage disequilibrium scores for mutations in *LuTFL1* and *LuTFL2*.

Electronic supplement 10: Raw data for correlation study in cultivated flax phenotypic traits.

Electronic supplement 11: Script written in *R* simulating the development of flax architecture in different latitudes with variable *FT*/*TFL1* expression.

Electronic supplement 12: The list of custom made P1 adapters used during RADseq library preparation.

Electronic supplement 13: The list of RADseq SNPs genotyped for flax in *STRUCTURE* format.

Electronic supplement 14: The *STRUCTURE* graphs constructed for number of K populations from 2 to 20.

Electronic supplement 15: The *INSTRUCT* graphs constructed for number of K populations from 2 to 20.

Electronic supplement 16: The output files of *MIGRATE-N* analyses containing marginal likelihoods.